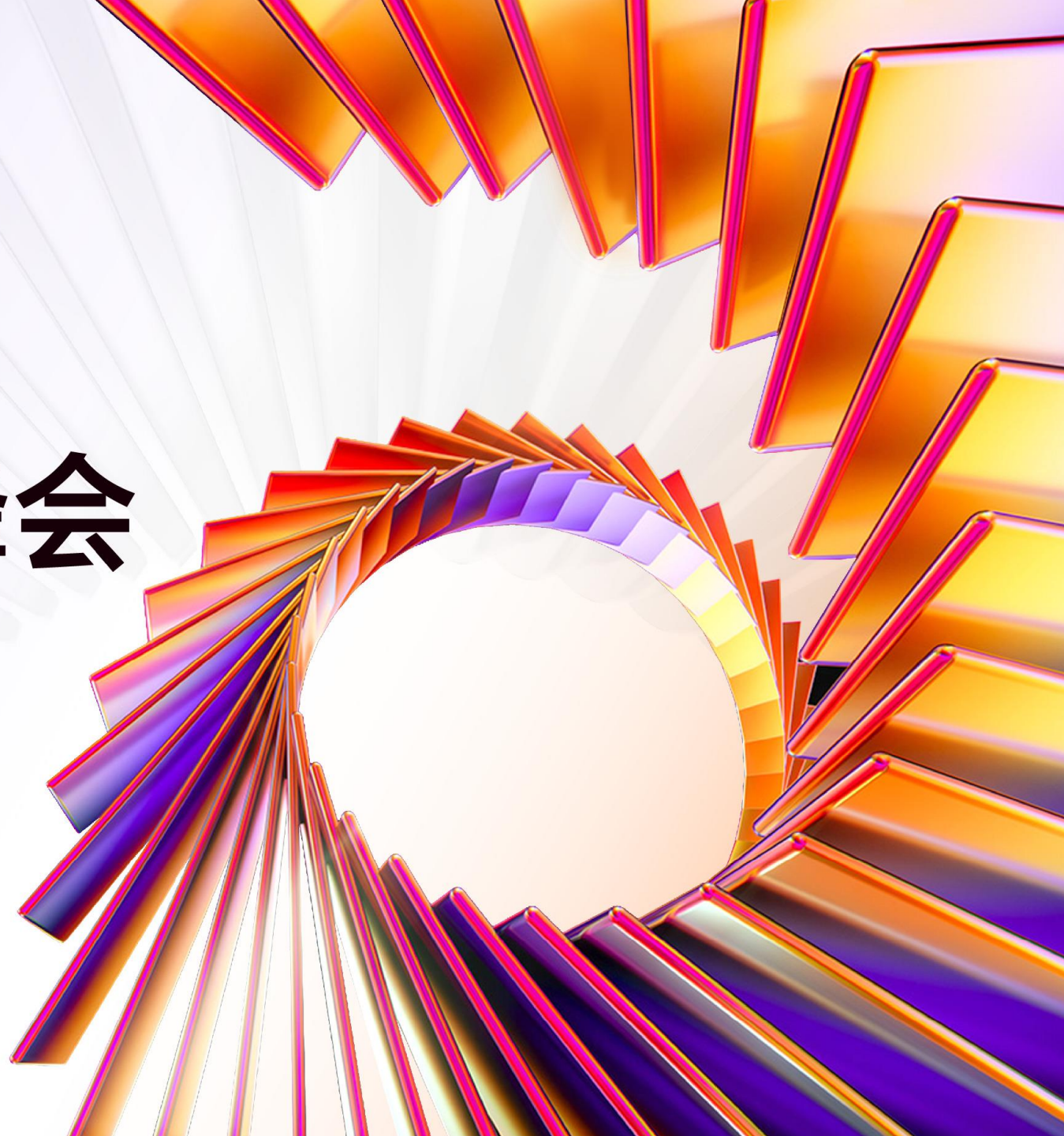
AiDD 6th 2025

2025 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发

05/23-24 | 上海站



AiDD 7th 2025

2025 AI+研发数字峰会

拥抱AI 重塑研发

AI+ Development Digital Summit

查看会议详情

下一站预告 08/08-09 | 北京站 11/14-15 | 深圳站

北京站论坛设置

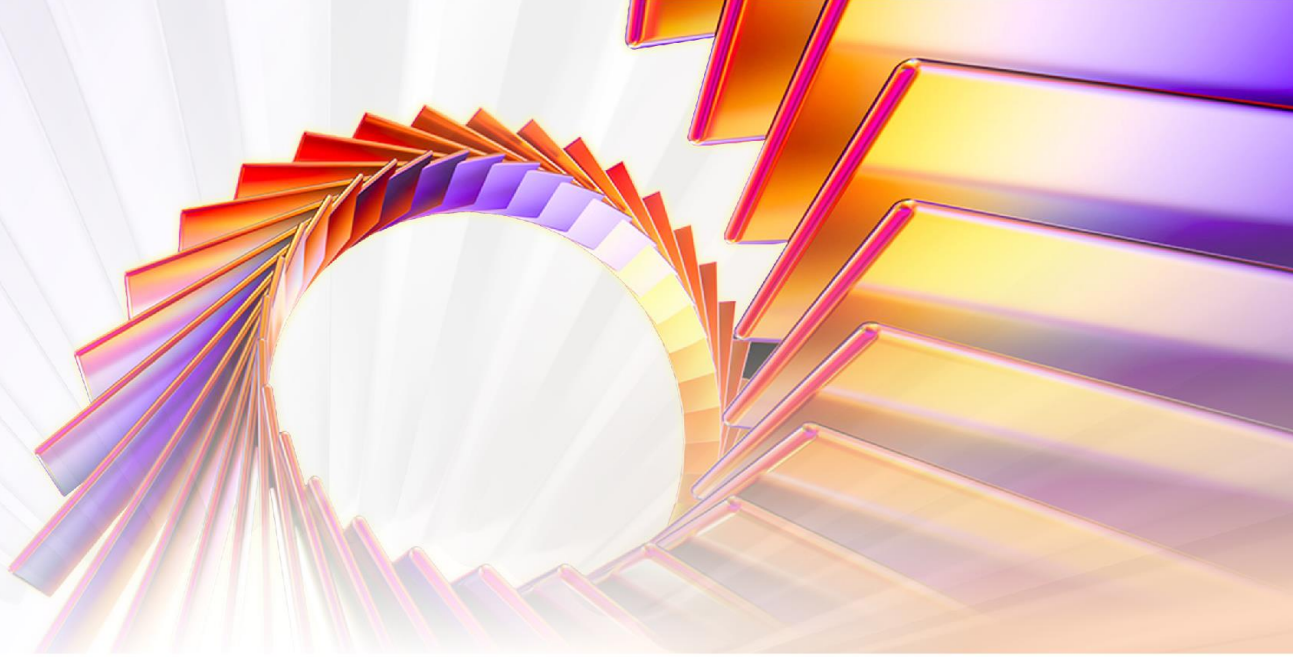
- AI+ 金融业务创新
- 大模型安全与对齐
- 具身智能与机器人
- 大模型和 AI 应用评测
- 智能需求工程
- 大模型应用开发框架与实践
- 代码生成及其改进
- 智能存储与检索技术
- 智能体与研发效率工具
- 智能体经济（Agentic Economy）
- AI+ 新能源汽车
- 下一代知识工程
- AI 产品运营与出海策略
- 智能测试工具的开发与应用
- AI 前沿技术探索与实践

AiDD 6th 2025 | 05/23-24 | 上海站

2025 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发



大小模型协同智能及端云协同应用

张圣宇 | 浙江大学

张圣宇

浙江大学平台”百人计划”研究员、博士生导师

浙江大学启真优秀青年学者。入选第十届中国科协青年人才托举工程。研究方向包括大小模型端云协同智能，多媒体计算与推荐系统。近年来，在TPAMI、TKDE、KDD、CVPR等CCF A类期刊和会议上发表论文四十余篇。任KDD、ACM MM 领域主席，SIGIR、IJCAI等会议高级程序委员会委员，ICMR、ACM MM Asia 大小模型协同workshop主席。曾获2023年度上海市科技进步一等奖、2023年度计算机学会科技进步一等奖，2024年ACM Multimedia最佳论文奖提名，2023年中国人工智能学会CICAI最佳论文奖、2021年WAIC云帆奖-明日之星（全球15人）等奖励与荣誉。

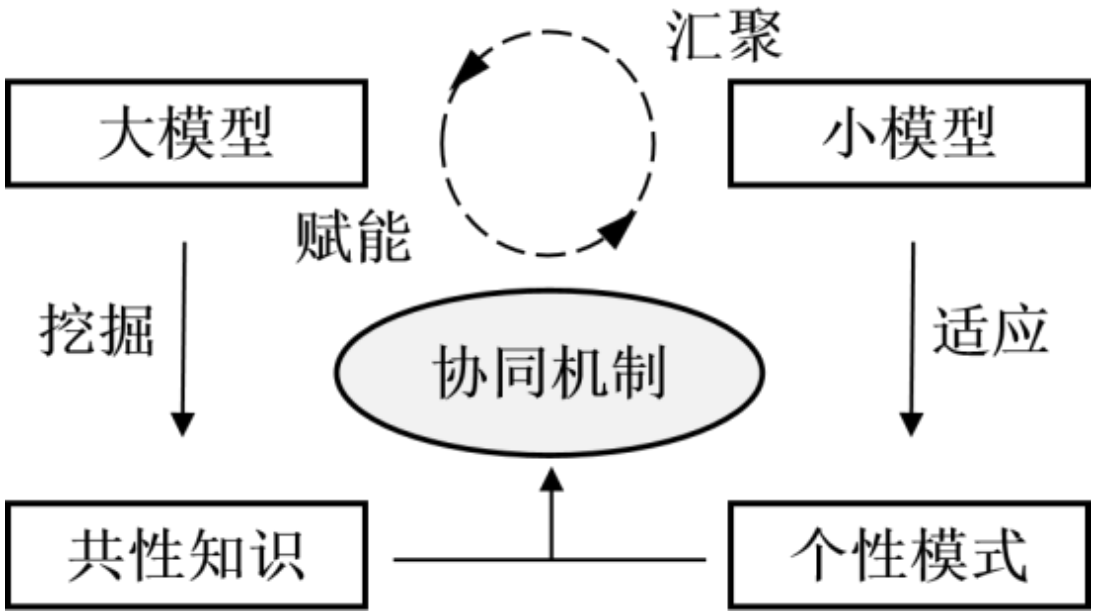
目录
CONTENTS

PART 01

大小模型端云协同智能的背景

大小模型端云协同智能

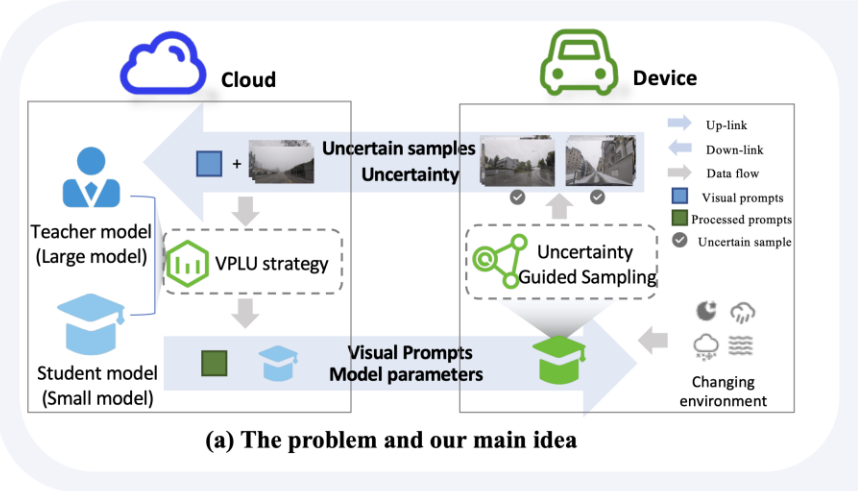
- 端云协同（Device-Cloud Collaboration）：指边缘设备（如智能手机、IoT设备）模型和云侧服务器模型协同进化推断。
- 云侧大模型（Large Model）：通用认知计算，拥有强大的计算能力、海量的数据、充分的知识库。
- 终端小模型（Small Model）：实时感知、实时响应，运行轻量级任务，响应速度快。



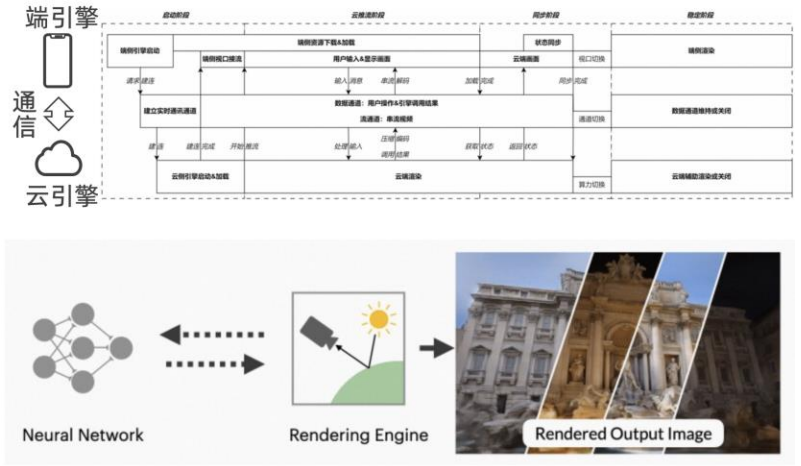
大小模型端云协同智能

- 端云协同计算通过卸载部分学习任务至端侧，让端和云协同完成任务，从而发挥终端靠近用户和数据源的天然优势，降低服务延时至毫秒级，增强模型个性化精准推理能力，缓解云服务器中心负载压力，同时支持用户原始数据在设备本地处理
- 有效克服主流云学习范式在实时性、个性化、负载成本、隐私安全等方面的不足

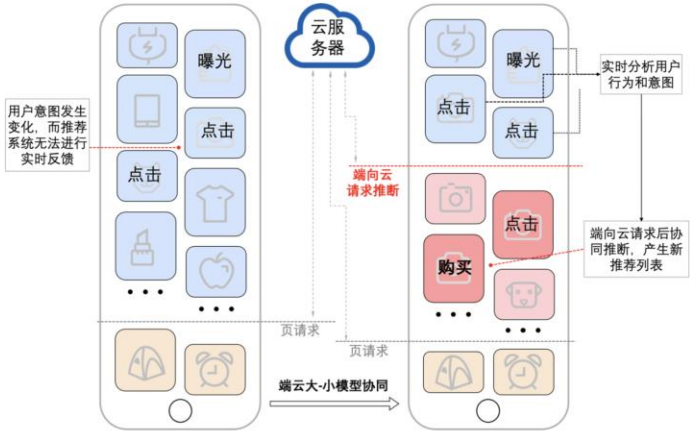
前沿应用



自动驾驶 (Gan et al.)



3D渲染 (Lv et al.)



推荐系统 (Qian et al.)

Yulu Gan, Mingjie Pan, Rongyu Zhang, et al.: Cloud-Device Collaborative Adaptation to Continual Changing Environments in the Real-World. CVPR 2023: 12157-12166
Chengfei Lv, Chaoyue Niu, Renjie Gu, et al.: Walle: An End-to-End, General-Purpose, and Large-Scale Production System for Device-Cloud Collaborative Machine Learning. OSDI 2022: 249-265
Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang, et al.: Intelligent Request Strategy Design in Recommender System. KDD 2022: 3772-3782

大小模型端云协同智能

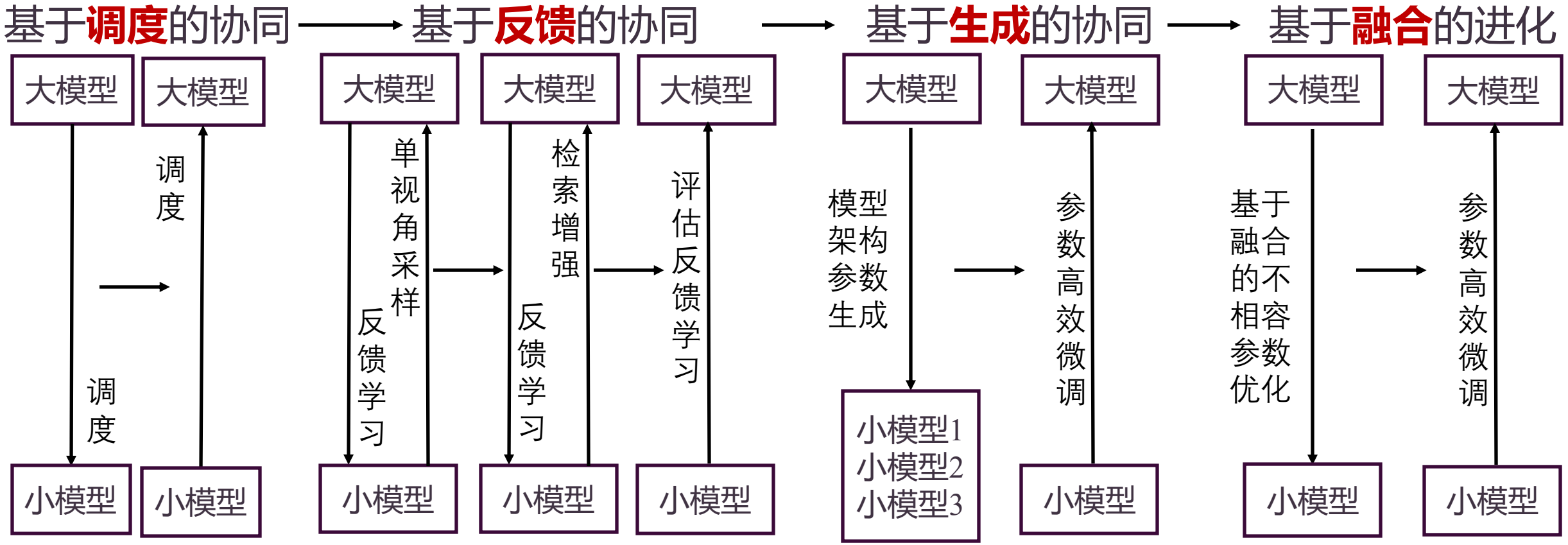

云
协同
链路
端
通用
大模型
泛化学习
轻量压缩
跨域迁移
自主请求
模型汇聚
高效适应
小模型
增强更新
轻量
小模型
小模型
实时推理

PART 02

大小模型协同基础算法

大小模型协同基础算法研究

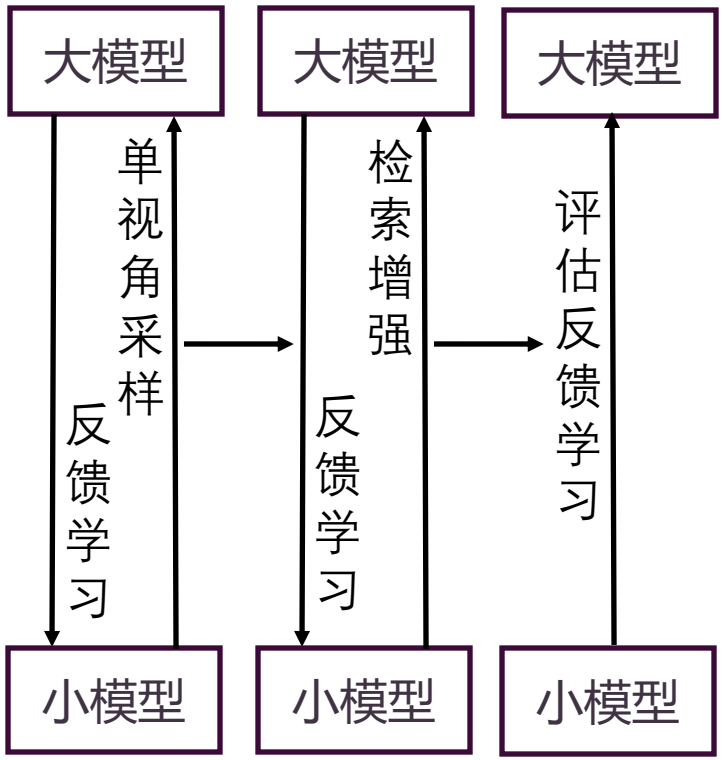
联合应用平台既有的特定业务小模型与云侧大模型，将端侧小模型轻量部署、快速响应、个性适配的优势，和云侧大模型认知推理、多模态理解、通用泛化的优势进行互补



ECCV	IntellectReq	MPOD123	Under	Work	ModelGPT	MergeNet	CKI	MoM
组合优化LLM	自主智能请求	2D到3D生成	Review	In Progress	Under Review	AAAI 25	AAAI 25	Under Review
ECCV 2024	WWW 24	CVPR 2024						

大小模型协同基础算法研究

基于调度的协同 → 基于反馈的协同 → 基于生成的协同 → 基于融合的进化



MPOD123
2D到3D生成
CVPR 2024

Under
Review

Work
In Progress

大语言模型强化学习综述

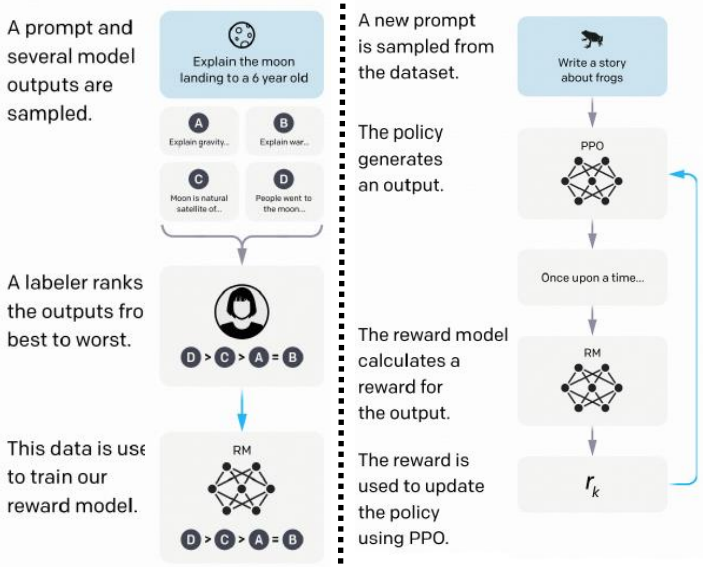
主要方法

- 基于人类反馈的强化学习（RLHF）：收集人类反馈数据，构造，进行模型优化
- 基于AI反馈的强化学习（RLAIF）：使用 AI 生成的数据代替人类反馈数据
- 直接偏好优化（DPO）：直接使用反馈数据优化模型

Reinforcement Learning Enhanced LLMs A Survey

Shuhe Wang, Shengyu Zhang, et al.
Under Review
(online, https://arxiv.org/pdf/2412.10400)

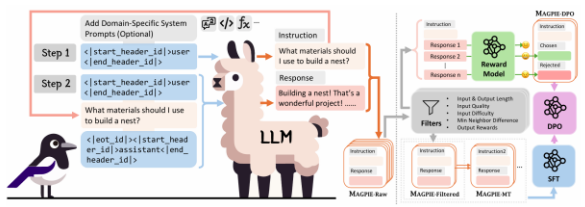
RHLF



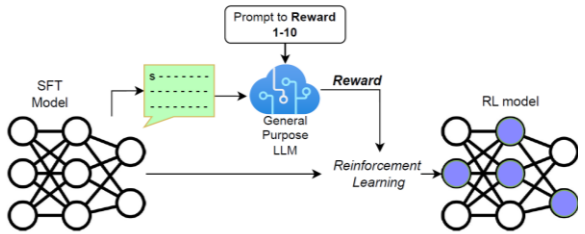
收集数据
构造奖励模型

PPO算法
训练模型

RLAIF



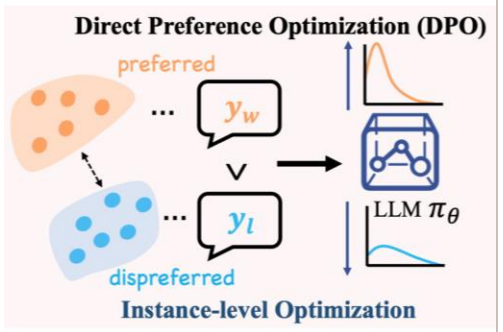
模型蒸馏代替人类反馈



（大）模型代替人类反馈

DPO

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\beta \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \Big[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \Big[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l} \Big] \Big]$$



收集数据
直接优化模型

Reinforcement Learning Enhanced LLMs: A Survey

Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, Eduard Hovy

arXiv:2412.10400v2 [cs.CL] 17 Dec 2024

Abstract

This paper surveys research in the rapidly growing field of enhancing large language models (LLMs) with reinforcement learning (RL), a technique that enables LLMs to improve their performance by receiving feedback in the form of rewards based on the quality of their outputs, allowing them to generate more accurate, coherent, and contextually appropriate responses. In this work, we make a systematic review of the most up-to-date state of knowledge on RL-enhanced LLMs, attempting to consolidate and analyze the rapidly growing research in this field, helping researchers understand the current challenges and advancements. Specifically, we (1) detail the basics of RL; (2) introduce popular RL-enhanced LLMs; (3) review researches on two widely-used reward model-based RL techniques: Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF); and (4) explore Direct Preference Optimization (DPO), a set of methods that bypass the reward model to directly use human preference data for aligning LLM outputs with human expectations. We will also point out current challenges and deficiencies of existing methods and suggest some avenues for further improvements. Project page of this work can be found at our latest repo.

1 Introduction

Large language models (Jiang et al., 2023; OpenAI, 2023; Dubey et al., 2024) are sophisticated language models pre-trained on extensive text data, allowing them to produce coherent and fluent responses to diverse inputs. However, the interaction capabilities of these pre-trained LLMs can be inconsistent, sometimes leading to responses that, while technically correct, may be harmful, biased, misleading, or irrelevant to users' needs. Therefore, it is crucial to align the outputs of pre-trained LLMs with human preferences before they can be effectively applied to various natural language tasks (Wang et al., 2023b; Wan et al., 2023; Sun et al., 2023c,b; Giray, 2023; Zhang, 2023; Long, 2023; Sun, 2023; Gao et al., 2023; Paranjape et al., 2023; Sun et al., 2023a; Diao et al., 2023; Wang et al., 2023a; Zhang et al., 2023b; Sun et al., 2023d; Liu et al., 2024d; Yao et al., 2024; Liu et al., 2024c; Lee et al., 2024; Kambhampati, 2024; Wang et al., 2024c).

Previously, a widely adopted approach for aligning the outputs of pre-trained LLMs with human preferences has been supervised fine-tuning (SFT) (Hu et al., 2021; Mishra et al., 2021; Wang et al., 2022; Du et al., 2022; Dettmers et al., 2023; Taori et al., 2023; Zhang et al., 2023a; Chiang et al., 2023; Xu et al., 2023; Peng et al., 2023; Mukherjee et al., 2023; Li et al., 2023; Ding et al., 2023; Luo et al., 2023; Wang et al., 2024d; Zhou et al., 2024). This method further trains LLMs on (Instruction, Answer) pairs, where "Instruction" represents the human prompt given to the model, and "Answer" is the target output that follows the instruction. SFT helps guide LLMs to produce responses that adhere to specific characteristics or domain knowledge, making it possible for humans to interactive with LLMs. Despite its effectiveness, SFT has limitations: during training, the model is constrained to learn specific answers we provide, with metrics like perplexity (PPL) penalizing synonym use. On one hand, this can hinder the LLM's ability to generalize, as tasks like writing and summarization have multiple valid phrasings. On the other hand, it may cause poor performance in aligning with human preferences, as no direct human feedback is incorporated into the training process.

The University of Melbourne, Zhejiang University, CFAR and IHPC, A*STAR, Singapore, Nanyang Technological University, University of Washington
Email: shuhewang@student.unimelb.edu.au

Project page of this work can be found at: https://github.com/ShuheWang1998/Reinforcement-Learning-Enhanced-LLMs-A-Survey

The latest update was on Dec. 3, 2024 (Version 1).

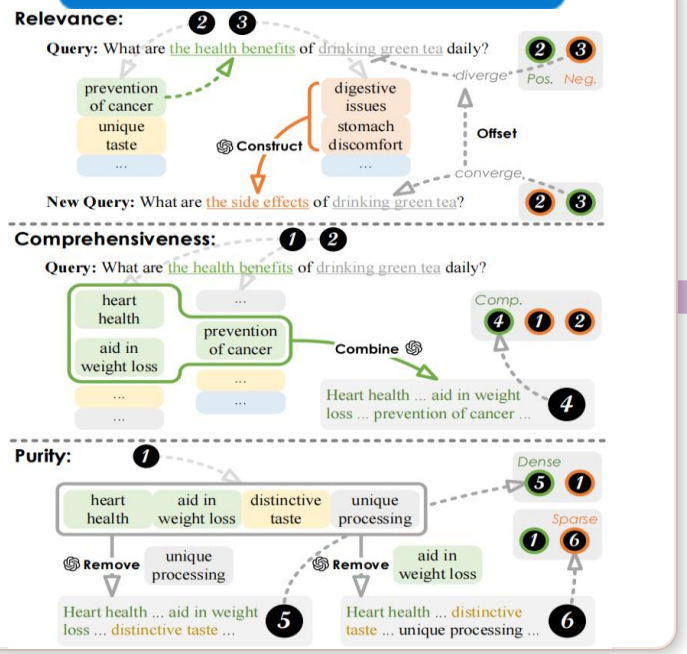
FiGRet: 基于 LLM 反馈的检索器细粒度优化

研究问题 LLM偏好信号复杂而传统密集检索器语义理解较弱，难以有效学习与对齐复杂偏好，影响RAG性能

解决方案

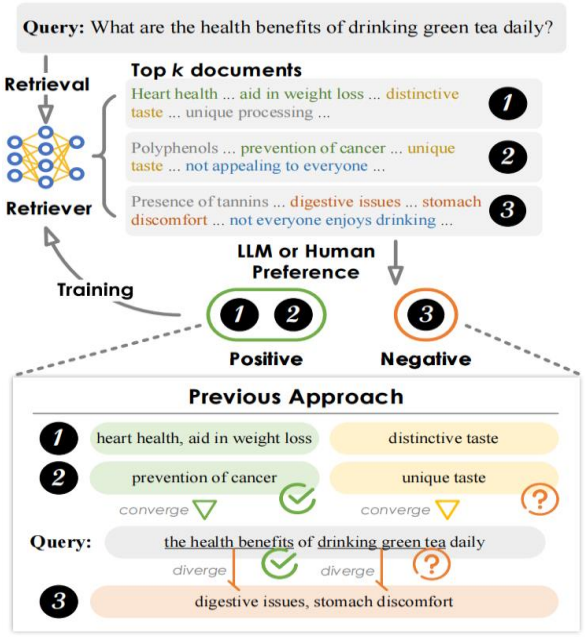
- 偏好认知注入，借鉴引导式学习，利用LLM提取高质量指导信息并注入检索器，帮助检索器建立LLM偏好认知
- 核心特征强化，聚焦相关性、全面性、纯净度三大RAG文档特征，结合双课程学习策略进一步提高检索器性能

LLM细粒度引导



支撑

检索增强偏好对齐



实验结果

Method	MMLU Hum.	MMLU Soc.	MMLU STEM	MMLU Other	MMLU All	NQ	PQA	HoPo	FEV.	All Avg.
GPT-3.5-Turbo										
No retrieval	52.9	76.6	53.1	75.7	63.4	48.1	44.3	33.6	82.1	54.3
Contriever	55.1	76.3	54.5	74.5	64.2	48.8	45.6	39.0	89.4	57.4
$AAR_{Contriever}$	54.3	78.5	52.5	77.1	64.4	49.0	46.3	36.9	89.6	57.2
BGE	52.9	78.2	54.0	76.5	64.1	50.3	43.9	39.5	89.3	57.4
SBERT	54.1	77.9	52.8	**77.4**	64.5	49.4	**50.1**	38.7	88.5	58.2
$FiGRet_{Contriever}$ (Ours)	55.4	76.9	54.5	77.1	65.0	49.6	48.0	39.9	**90.6**	**58.6**
$FiGRet_{BGE}$ (Ours)	**55.8**	**79.8**	54.3	76.5	**65.5**	**50.4**	45.7	**40.0**	90.3	58.4
$FiGRet_{SBERT}$ (Ours)	**55.8**	77.2	**55.2**	76.8	65.4	49.9	**50.1**	39.1	88.7	**58.6**
Llama-3-8B-Instruct										
No retrieval	52.9	74.4	51.6	73.3	62.0	33.1	26.1	25.9	79.1	45.2
Contriever	52.9	76.3	52.5	73.3	62.5	41.3	41.7	36.0	84.5	53.2
$AAR_{Contriever}$	52.9	77.2	54.0	73.9	63.2	42.1	42.3	35.3	85.2	53.6
BGE	54.4	76.9	52.8	73.6	63.3	44.1	36.1	35.9	86.1	53.1
SBERT	53.9	76.6	**54.6**	73.3	63.4	41.7	46.0	35.7	86.2	54.6
$FiGRet_{Contriever}$ (Ours)	53.5	**77.5**	53.1	74.8	63.4	43.0	44.3	36.9	86.5	54.8
$FiGRet_{BGE}$ (Ours)	53.9	76.3	53.4	**75.1**	63.6	**45.3**	41.4	**37.5**	**87.8**	**55.1**
$FiGRet_{SBERT}$ (Ours)	**54.6**	76.3	54.0	74.2	**63.7**	42.8	**46.3**	36.1	86.2	55.0
Claude-3-Haiku										
No retrieval	59.5	82.6	59.4	78.3	68.8	27.6	31.7	26.9	70.4	45.1
Contriever	61.6	82.0	60.0	79.8	70.0	35.7	41.3	33.0	90.0	54.0
$AAR_{Contriever}$	62.6	83.5	59.1	79.5	70.2	36.1	42.1	32.7	90.2	54.3
BGE	61.4	82.0	58.5	77.4	69.0	38.1	37.5	33.0	89.6	53.4
SBERT	62.0	81.0	58.8	79.2	69.4	35.9	46.2	32.7	89.5	54.7
$FiGRet_{Contriever}$ (Ours)	62.9	82.0	**60.3**	**80.1**	**70.5**	36.5	44.2	33.8	**90.4**	55.1
$FiGRet_{BGE}$ (Ours)	63.3	82.9	57.9	78.9	69.9	**40.0**	42.2	**35.7**	90.0	**55.6**
$FiGRet_{SBERT}$ (Ours)	**63.7**	**84.2**	59.4	77.7	**70.5**	37.1	**46.6**	33.0	90.1	55.5

跨LLM，跨Base Retriever均取得性能提升

FiGRet能增强各类检索器与各系列LLM在多种任务上的协作效果，同时实现样本高效学习

基于反馈的协同：大模型反馈小模型

思路方法

- Stage 1

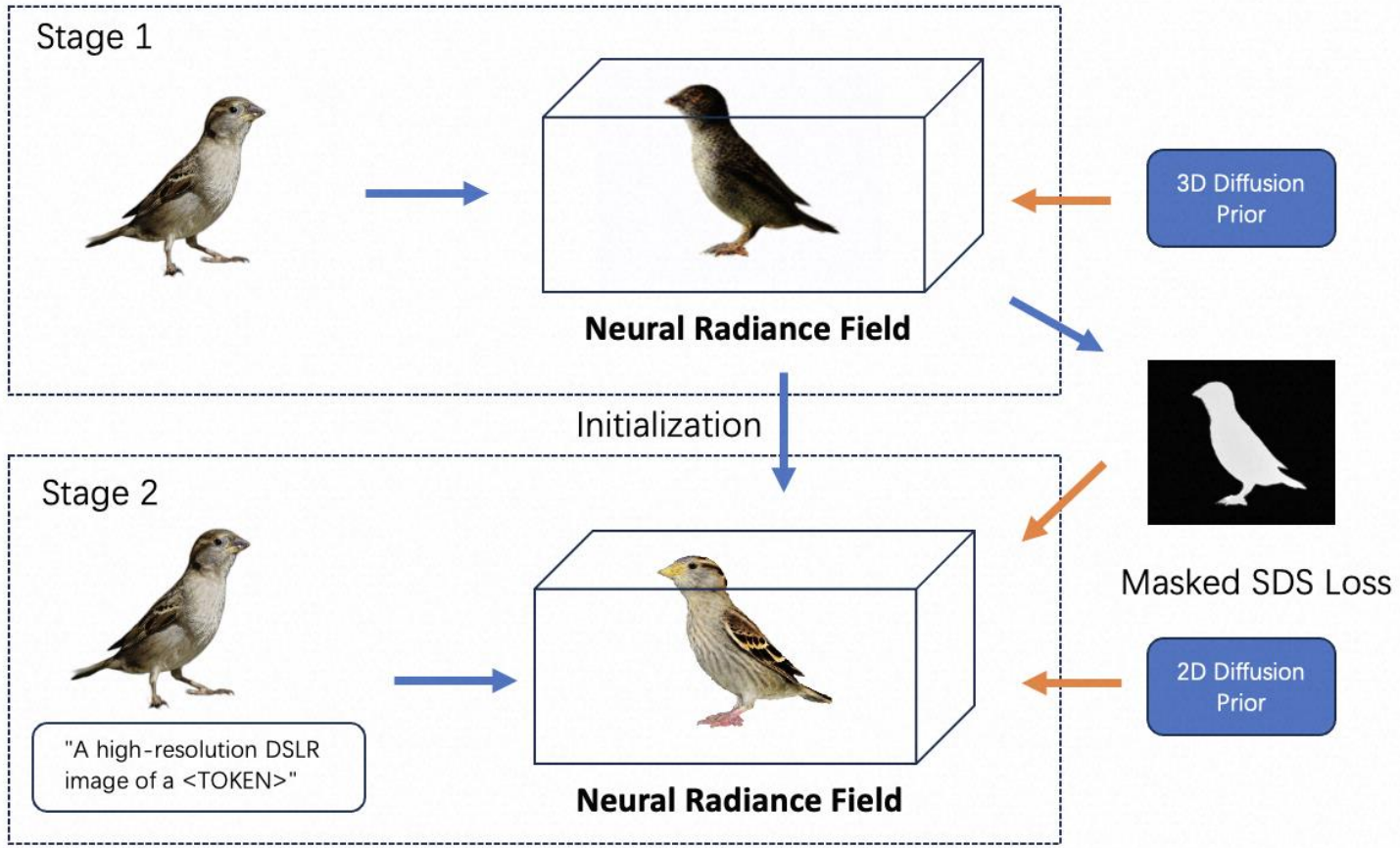
View-Conditioned Diffusion Priors

对于给定视角，能生成较好的形状，但在纹理等细节上质量差。

- Stage 2

Diffusion Inpainting Priors

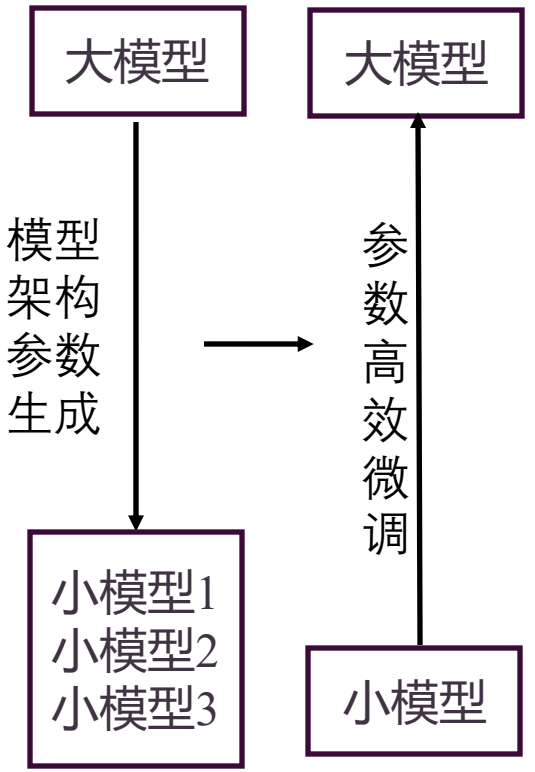
基于Stage 1的NeRF模型生成Mask，重绘Mask部分。在保持Stage 1形状的同时生成较好的纹理等细节。



MPOD123: One Image to 3D Content Generation Using Mask-enhanced Progressive Outline-to-Detail Optimization. CVPR 2024

大小模型协同基础算法研究

基于调度的协同 → 基于反馈的协同 → 基于生成的协同 → 基于融合的进化



ECCV 组合优化LLM ECCV 2024

IntellectReq 自主智能请求 WWW 24

MPOD123 2D到3D生成 CVPR 2024

Under Review

Work In Progress

ModelGPT Under Review

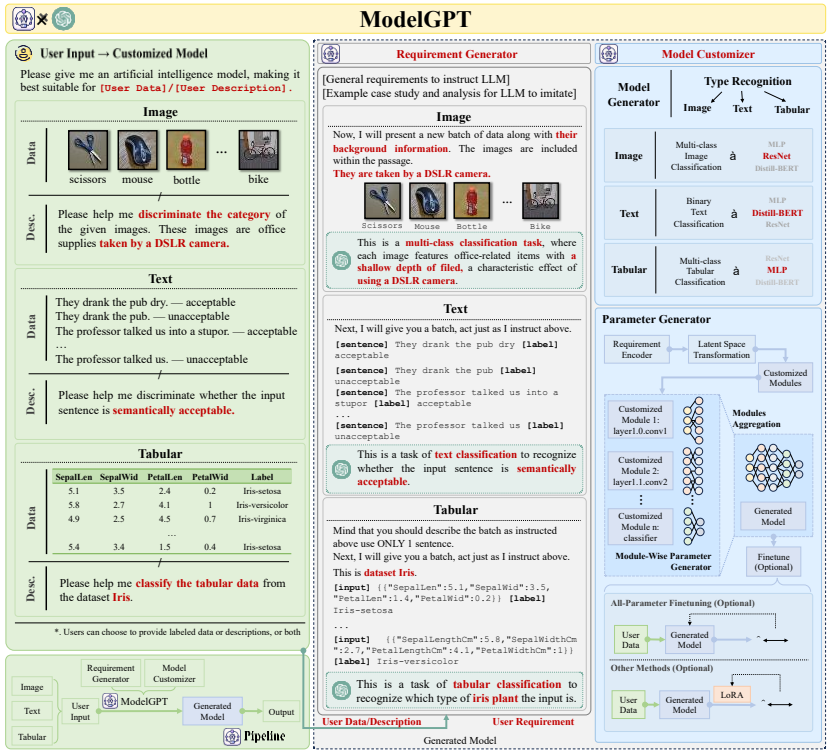
MergeNet AAAI 25

CKI AAAI 25

MoM Under Review

基于生成的协同：One（大模型） to All （小模型）生成

- **大模型驱动的小模型生成框架ModelGPT**
  - ModelGPT + 用户对**模型的需求描述** + **少量数据** =(推理生成) 开箱即用小模型。在All-in-One 的通用大模型范式之外，初步探索 **One-to-All** 的可能性，为更广泛的小数据、小算力（边端）、离线应用场景提供AI落地支撑。
  - 在**NLP，CV，和Tabular Data**典型数据集上进行验证，**性能超越Finetune**方法。



Algorithm 1 Pseudo-code of Parameter Generator $P(\cdot;\theta_p)$

Require: $A = \{(D_i = \{X_i, Y_i\}, r_i)\}_{i=1}^{N}$
Ensure: $\theta_p = (\theta_e, \theta_m, \theta_g)$ satisfies Equation (5)

```
i ← 1
for _ = 0 to ♯epoch do
  for (D_i, r_i) in A do
    for batch in D_i do
      Obtain θ_t with Equations (1) to (3)
      Use batch to compute the loss and update θ_t
      Compute the difference Δθ_t of θ_t
      Use Δθ_t to compute the gradients of θ_p
      Update θ_p
    end for
  end for
  Save best checkpoint according to Equation (5)
end for
```

Zihao Tang, Zheqi Lv, Shengyu Zhang, Fei Wu, Kun Kuang: ModelGPT: Unleashing LLM's Capabilities for Tailored Model Generation. CoRR abs/2402.12408 (2024)

基于生成的协同：One（大模型） to All （小模型）生成

- **大模型驱动的小模型生成框架ModelGPT**
  - 在**NLP，CV，和Tabular Data**典型数据集上进行验证，**性能超越Finetune**方法。
  - 给定用户的需求ModelGPT能够以至多先前范式（例如全参数微调、LORA微调）**270倍速度**快速生成定制好的人工智能模型。

Results on GLUE Benchmark (Distil-BERT)

Methods	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	DM	Score	♯Epoch	E2E Runtime
LoRA	**48.3**	91.0	84.9 / 80.3	81.2 / 80.0	68.9 / 87.3	80.5	33.1	**88.1**	52.8	**65.1**	0.0	71.5	20	216.1
Finetune	45.5	**91.3**	**86.6 / 80.8**	**82.1 / 80.9**	**69.2 / 87.8**	**81.8**	**80.8**	87.6	56.9	63.7	**35.6**	**74.4**	20	273.8
ModelGPT	39.5	88.9	85.3 / 78.4	80.9 / 80.3	63.3 / 83.5	77.8	78.0	84.6	69.5	64.4	28.0	73.4	**0**	**1.0**
ModelGPT-F	36.9	90.8	85.5 / 79.4	81.3 / 80.5	67.0 / 86.6	77.8	78.1	85.8	**70.0**	62.3	29.9	73.8	1	3.1

Results on Tabular Data (MLP)

Methods	Iris	Heart Disease	Wine	Adult	Breast Cancer	Car Evaluation	Wine Quality	Dry Bean	Rice	Bank Marketing	Average	♯Epoch	E2E Runtime
LoRA	93.3	**63.0**	67.3	54.7	95.9	71.3	55.0	88.9	92.5	89.8	77.2	20	46.2
Finetune	88.9	54.3	89.1	**55.2**	**96.5**	71.0	55.3	**90.6**	**93.1**	89.9	78.4	20	39.5
ModelGPT	**100.0**	60.9	**94.5**	54.7	95.3	**71.5**	54.1	85.0	92.5	89.8	79.8	**0**	**1.0**
ModelGPT-F	**100.0**	62.0	**94.5**	55.1	95.9	71.3	**55.4**	88.8	92.9	**90.0**	**80.6**	1	2.3

Results on Office-31 (ResNet-50)

Domain	Amazon			DSLR			Average			Webcam (ModelGPT is Zero-Shot)			♯Epoch	E2E Runtime
Methods	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5		
LoRA	66.4	77.7	84.8	78.4	92.2	96.1	72.4	85.0	90.5	72.5	87.5	93.8	400	231.8
Finetune	67.5	79.2	83.7	84.3	98.0	**100.0**	75.9	88.6	91.9	**90.0**	**100.0**	**100.0**	400	257.6
ModelGPT	66.4	79.9	83.7	**92.2**	**100.0**	**100.0**	79.3	90.0	91.9	76.2	87.5	91.2	**0**	**1.0**
ModelGPT-F	**67.8**	**81.3**	**85.9**	**92.2**	**100.0**	**100.0**	**80.0**	**90.7**	**92.8**	77.5	90.0	91.3	1	1.2

跨越异构模型、任务、模态的统一模型知识迁移框架

研究背景

现有知识迁移方法（例如，知识蒸馏，迁移学习）要求端云具有相似的任务类型或模型架构，难以应用于**跨异构模型、任务和模态**的异构知识迁移场景。

挑战

模型知识统一表示

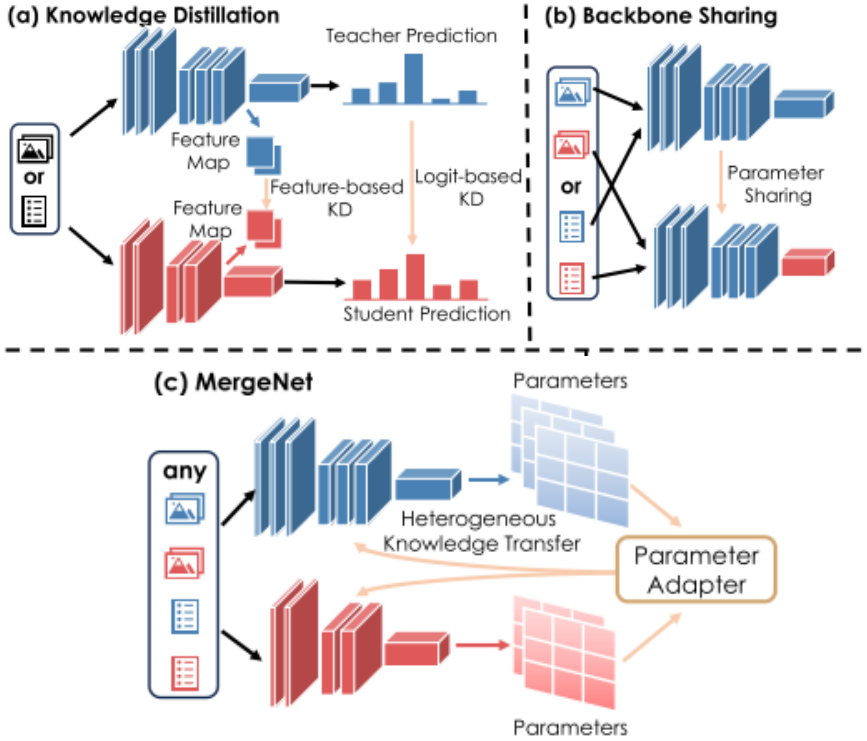
知识蒸馏利用**Logits**和**Feature Map**表示知识，依赖于任务类型。

迁移学习通常通过**共享参数**实现知识迁移，依赖于模型架构。

异构模型知识适配

异构模块（线性层 <-> 注意力机制模块）之间知识不兼容。

不同规模模型之间知识不兼容。



Li K, Zhan T, Fu K, Zhang S, et al. MergeNet: Knowledge Migration across Heterogeneous Models, Tasks, and Modalities. AAAI 2025

跨越异构模型、任务、模态的统一模型知识迁移框架

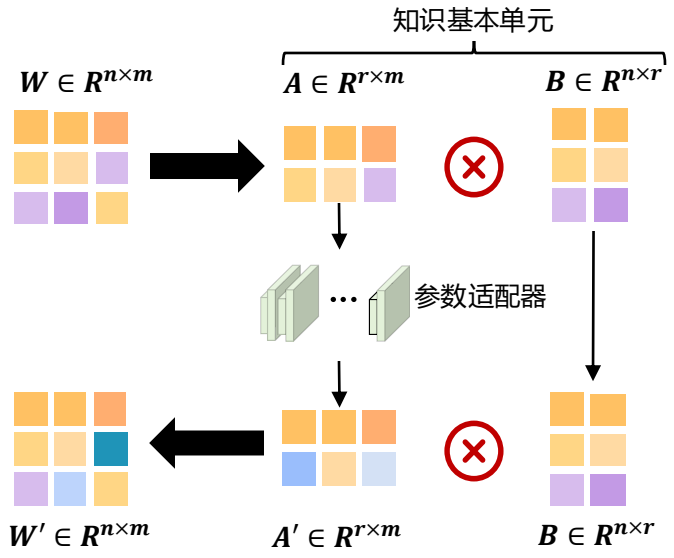
研究问题　研究基于端云协同的跨异构模型架构、任务和模态的异构知识迁移框架。

创新方法

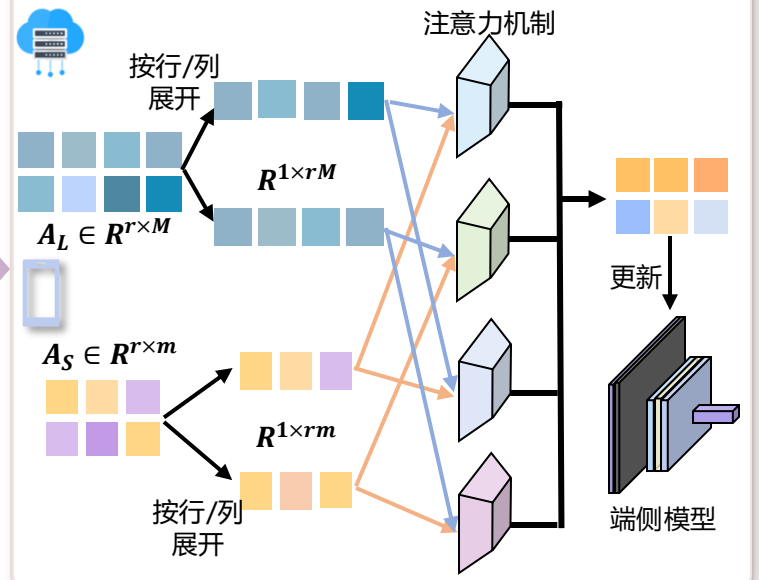
异构模型知识表示：以参数为载体，重新编码端云模型参数，实现对异构知识的统一表示

异构知识适配：设立参数适配器，促进异构参数空间的交互，提取并对齐有效的信息，实现高效知识迁移

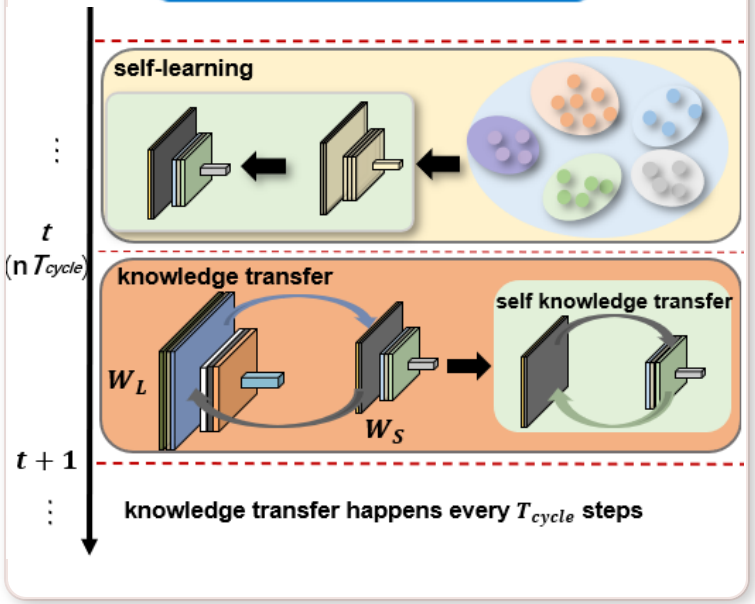
异构模型知识表示



异构知识适配



协同训练流程



Li K, Zhan T, Fu K, Zhang S, et al. MergeNet: Knowledge Migration across Heterogeneous Models, Tasks, and Modalities. AAAI 2025

跨越异构模型、任务、模态的统一模型知识迁移框架

应用验证

克服了传统知识迁移需要具有相似任务类型或模型架构的限制

有效应用于各种具有挑战性的场景，及传统知识迁移方法不适用的场景

传统知识迁移存在的问题

模型结构差异性限制
任务类型匹配要求
异构知识表示不兼容

统一异构知识表示
知识交互融合
…

异构知识迁移

跨架构知识迁移

Methods	Top-1 Acc(%)	Top-5 Acc(%)
Vanilla MobileNetV2	63.87	88.77
KD (Hinton et al., 2015)	64.32	88.62
RKD (Park et al., 2019)	65.48	88.9
DKD (Zhao et al., 2022)	65.23	89.01
NKD (Yang et al., 2023)	65.09	88.9
MergeNet(R→M)	66.23	89.66
MergeNet(R↔M)	**66.51**	**89.75**
Vanilla ResNet50	68.11	89.61
KD(Hinton et al., 2015)	68.36	89.9
RKD(Park et al., 2019)	68.6	90.21
DKD (Zhao et al., 2022)	69.03	90.25
NKD(Yang et al., 2023)	69.27	90.18
MergeNet(R↔M)	**69.84**	**90.57**

跨模态知识迁移

Methods	VQA overall	VQA other	VQA number	ITR TR	ITR IR
Vanilla	45.78	31.33	28.71	41.48	37.64
MergeNet(V→T)	46.33	33.29	31.33	44.72	39
MergeNet(T→V)	45.96	31.99	31.15	44.58	38.93
MergeNet	**46.51**	**33.84**	**31.54**	**44.78**	**39.26**

跨任务知识迁移

Methods	SQuAD v2.0 EM↑	SQuAD v2.0 F1↑	IMDb Err↓
Vanilla	70.17	73.06	8.02
MergeNet	71.89	75.43	7.5

Methods	CIFAR-100 Top-1 Acc	CIFAR-100 Top-5 Acc	SQuAD v2.0 EM	SQuAD v2.0 F1
Vanilla	63.87	88.77	70.17	73.06
MergeNet	65.56	88.74	70.89	74.15

Li K, Zhan T, Fu K, Zhang S, et al. MergeNet: Knowledge Migration across Heterogeneous Models, Tasks, and Modalities. AAAI 2025

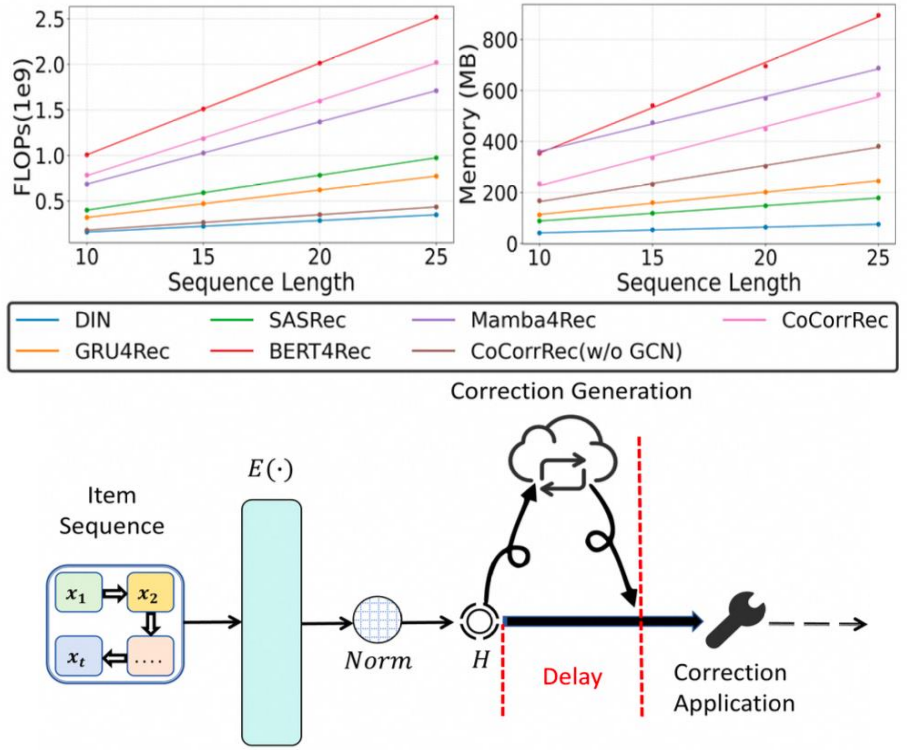
端云协同测试时学习（Collaborative Learning to Learn at Test Time）

研究背景

随着端侧计算能力的提升，部署在端侧的推荐系统因具备实时性和隐私保护优势而成为重要研究方向，但现有模型在资源受限环境下的计算开销仍存在挑战。

资源消耗与性能平衡

- 测试时训练：Test-Time Training (TTT) 是一种在测试阶段动态更新模型参数的技术，主要用于提升模型的性能，同时保持低计算开销。其核心思想是通过自监督学习和mini-batch加速策略，在实时推理过程中调整模型参数。**但在端侧持续更新参数会存在过拟合问题**。
- 协同过滤：在端侧推荐系统中引入协同过滤信息能够有效缓解模型过拟合风险，同时保留模型对全局数据特征的学习能力。然而，该信息的提取需要**消耗显著的计算资源**，若直接部署在端侧，会导致**服务响应延迟**。



Zhan Tianyu, Zhang Shengyu, Lv Zheqi, et al. Device-Cloud Collaborative Correction for On-Device Recommendation. IJCAI 2025

端云协同测试时学习（Collaborative Learning to Learn at Test Time）

研究问题

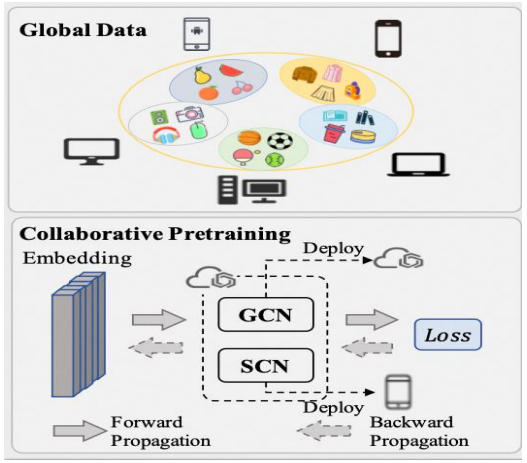
部署端云协同参数校正框架，保证端侧低计算开销和实时性的同时，提升推荐任务的性能。

创新方法

测试时训练：通过自监督学习在推理阶段动态更新模型参数，利用低资源开销实现推荐性能优化。

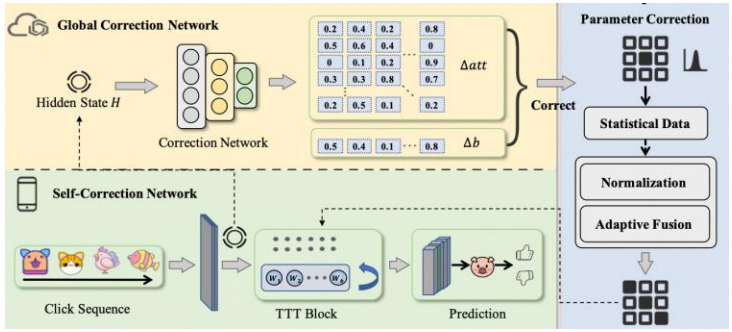
云侧参数校正：基于云端全局用户行为数据提取共性特征，生成校正参数维持端侧模型能力避免过拟合。

协同训练



在云侧基于收集到的全局数据进行协同训练

协同推荐



端侧根据用户兴趣实时更新参数，云侧基于全局数据提供参数校正避免过拟合

模型性能提升

Dataset	Model	Metrics						
		UAUC	NDCG@5	NDCG@10	NDCG@20	HitRate@5	HitRate@10	HitRate@20
Beauty	DIN	0.7185	0.2525	0.2866	0.3102	0.3655	0.4703	0.5629
	GRU4Rec	0.7078	0.2577	0.2825	0.3103	0.3485	0.4254	0.5361
	SASRec	0.7314	0.2605	0.2920	0.3180	0.3929	0.4852	0.5876
	BERT4Rec	0.7178	0.2555	0.2829	0.3109	0.3445	0.4296	0.5403
	Mamba4Rec	0.7220	0.3051	0.3300	0.3543	0.3871	0.4639	0.5603
	CoCorrRec (w/o. GCN)	0.7331	0.2987	0.3256	0.3517	0.4019	0.4858	0.5874
	CoCorrRec	**0.7342**	**0.3126**	**0.3371**	**0.3626**	**0.4108**	**0.4867**	**0.5888**
Electronic	DIN	0.7991	0.3422	0.3799	0.4070	0.4596	0.5760	0.6829
	GRU4Rec	0.8257	0.3528	0.3933	0.4237	0.4762	0.6016	0.7218
	SASRec	0.8161	0.3418	0.3820	0.4131	0.4639	0.5881	0.7107
	BERT4Rec	0.8182	0.3454	0.3851	0.4171	0.4665	0.5891	0.7156
	Mamba4Rec	0.8198	0.3390	0.3787	0.4115	0.4632	0.5861	0.7157
	CoCorrRec (w/o. GCN)	0.8258	0.3550	0.3950	0.4258	0.4754	0.5989	0.7210
	CoCorrRec	**0.8315**	**0.3629**	**0.4021**	**0.4327**	**0.4862**	**0.6074**	**0.7282**
Yelp	DIN	0.9522	0.5302	0.5774	0.5959	0.7333	0.8775	0.9497
	GRU4Rec	0.9522	0.5247	0.5724	0.5916	0.7318	0.8778	0.9525
	SASRec	0.9499	0.5099	0.5607	0.5804	0.7173	0.8729	0.9495
	BERT4Rec	0.9524	0.5388	0.5830	0.6000	0.7451	0.8841	0.9514
	Mamba4Rec	0.9514	0.5301	0.5776	0.5954	0.7387	0.8832	0.9528
	CoCorrRec (w/o. GCN)	0.9511	0.5372	0.5835	0.6006	0.7412	0.8824	0.9489
	CoCorrRec	**0.9533**	**0.5412**	**0.5865**	**0.6036**	**0.7470**	**0.8858**	**0.9531**

相较于常用的端侧推荐模型，在降低端侧负载的同时，实现推荐性能的提升

端云协同测试时学习（Collaborative Learning to Learn at Test Time）

应用验证

当前端云推荐
常见的问题
通信开销大——传递中间状态
云端分布差异大——云侧提供参数校正
端侧兴趣变化快——推理时更新参数
设备计算资源有限——采用TTT减少开销

实现了端云协同计算在**分布偏移、资源受限**设备上性能提升同时降低负载

云侧校正参数下载所需时间在执行校正之前，不会引入额外延迟

Size	4G: 5MB/s	4G: 15MB/s	5G: 50MB/s	5G: 100MB/s	Tolerance
⇑: 2.56KB ○ : 1.18ms ⇓: 4.16KB	⇑: 0.51ms ○ : 1.18ms→ 2.52ms ⇓: 0.83ms	⇑: 0.17ms ○ : 1.18ms→ 1.63ms ⇓: 0.28ms	⇑: 0.051ms ○ : 1.18ms → 1.32ms ⇓: 0.083ms	⇑: 0.026ms ○ : 1.18ms → 1.25ms ⇓: 0.042ms	4.21ms

端侧所需计算资源减少，缓和资源受限环境下的困难

Model	#Parameter	△FLOPs	△Memory
DIN	1.97M	12.6M	2.25M
GRU4Rec	1.97M	30.2M	8.78M
SASRec	3.91M	38.3M	6.01M
BERT4Rec	2.02M	101.0M	35.6M
Mamba4Rec	2.06M	68.4M	21.6M
CoCorrRec (W/o GCN)	1.95M	17.0M	14.2M
CoCorrRec	2.08M	82.7M	23.2M

在多个数据集上实现性能提升（0.06%-4.56%）

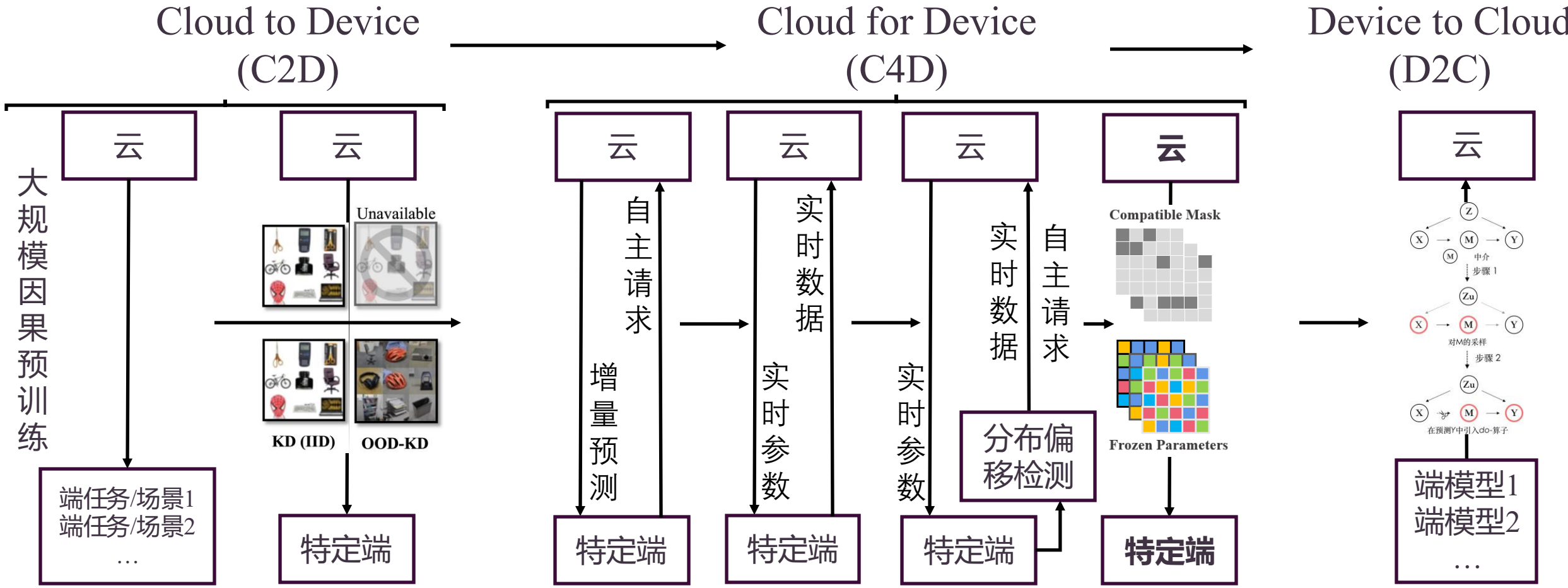
Dataset	Model	Metrics						
		UAUC	NDCG@5	NDCG@10	NDCG@20	HitRate@5	HitRate@10	HitRate@20
Beauty	DIN	0.7185	0.2525	0.2866	0.3102	0.3655	0.4703	0.5629
	GRU4Rec	0.7078	0.2577	0.2825	0.3103	0.3485	0.4254	0.5361
	SASRec	0.7314	0.2605	0.2920	0.3180	0.3929	0.4852	0.5876
	BERT4Rec	0.7178	0.2555	0.2829	0.3109	0.3445	0.4296	0.5403
	Mamba4Rec	0.7220	0.3051	0.3300	0.3543	0.3871	0.4639	0.5603
	CoCorrRec (w/o. GCN)	0.7331	0.2987	0.3256	0.3517	0.4019	0.4858	0.5874
	CoCorrRec	**0.7342**	**0.3126**	**0.3371**	**0.3626**	**0.4108**	**0.4867**	**0.5888**
Electronic	DIN	0.7991	0.3422	0.3799	0.4070	0.4596	0.5760	0.6829
	GRU4Rec	0.8257	0.3528	0.3933	0.4237	0.4762	0.6016	0.7218
	SASRec	0.8161	0.3418	0.3820	0.4131	0.4639	0.5881	0.7107
	BERT4Rec	0.8182	0.3454	0.3851	0.4171	0.4665	0.5891	0.7156
	Mamba4Rec	0.8198	0.3390	0.3787	0.4115	0.4632	0.5861	0.7157
	CoCorrRec (w/o. GCN)	0.8258	0.3550	0.3950	0.4258	0.4754	0.5989	0.7210
	CoCorrRec	**0.8315**	**0.3629**	**0.4021**	**0.4327**	**0.4862**	**0.6074**	**0.7282**
Yelp	DIN	0.9522	0.5302	0.5774	0.5959	0.7333	0.8775	0.9497
	GRU4Rec	0.9522	0.5247	0.5724	0.5916	0.7318	0.8778	0.9525
	SASRec	0.9499	0.5099	0.5607	0.5804	0.7173	0.8729	0.9495
	BERT4Rec	0.9524	0.5388	0.5830	0.6000	0.7451	0.8841	0.9514
	Mamba4Rec	0.9514	0.5301	0.5776	0.5954	0.7387	0.8832	0.9528
	CoCorrRec (w/o. GCN)	0.9511	0.5372	0.5835	0.6006	0.7412	0.8824	0.9489
	CoCorrRec	**0.9533**	**0.5412**	**0.5865**	**0.6036**	**0.7470**	**0.8858**	**0.9531**

Zhan Tianyu, Zhang Shengyu, Lv Zheqi, et al. Device-Cloud Collaborative Correction for On-Device Recommendation. IJCAI 2025

PART 03

大小模型端云协同智能

端云异构模型知识互迁与协同推断



DeVLBert/DeVADG
跨任务/场景泛化
ACM MM 20/AAAI 23

AUG-KD
迁移压缩
ICLR 24

AdaRequest
自主请求
KDD 22

DUET
实时适应
WWW 23

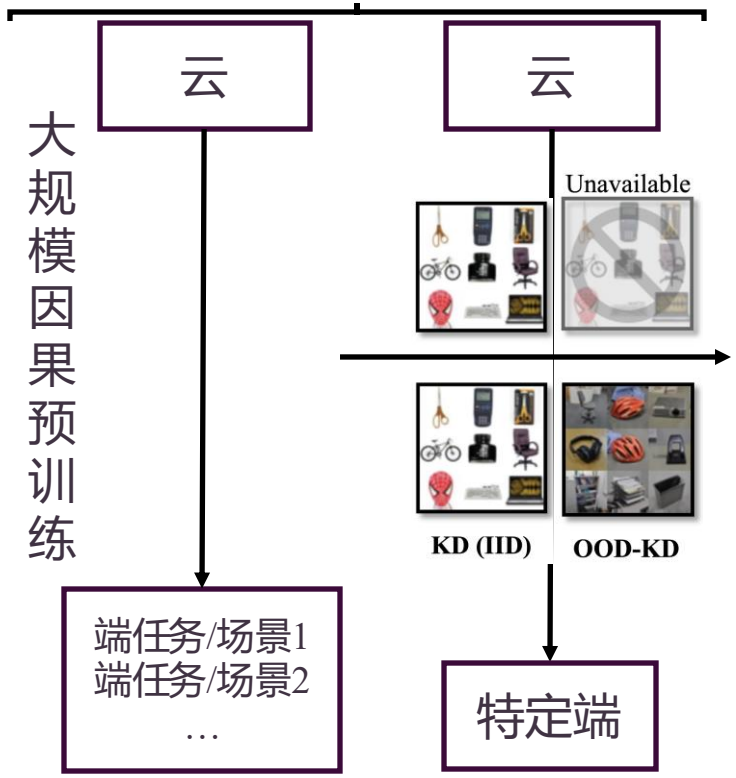
IntellectReq
实时自主适应
WWW 24

DIET/
Forward-OFA
高效定制
KDD 24/KDD 25

FedCFA/CausalD
因果去偏汇聚
AAAI 25, TKDE 23

端云异构模型知识互迁与协同推断

Cloud to Device
(C2D)



DeVLBert/DeVADG
跨任务/场景泛化
ACM MM 20/AAAI 23

AUG-KD
迁移压缩
ICLR 24

不同端设备存在**差异化任务功能**和**差异化使用场景**，云模型向端侧迁移部署面临着**跨场景、跨任务的泛化性**问题

面向未知端侧分布的压缩-适应联合

研究背景

- 大模型向端侧迁移部署往往采用知识蒸馏等压缩手段，传统知识整理方法假设大模型训练数据分布（压缩前）和小模型测试数据分布（压缩后）服从独立同分布假设（IID Hypothesis）。
- 实际应用中，源域数据和应用场景存在**分布偏移**，**导致压缩性能显著下降。**

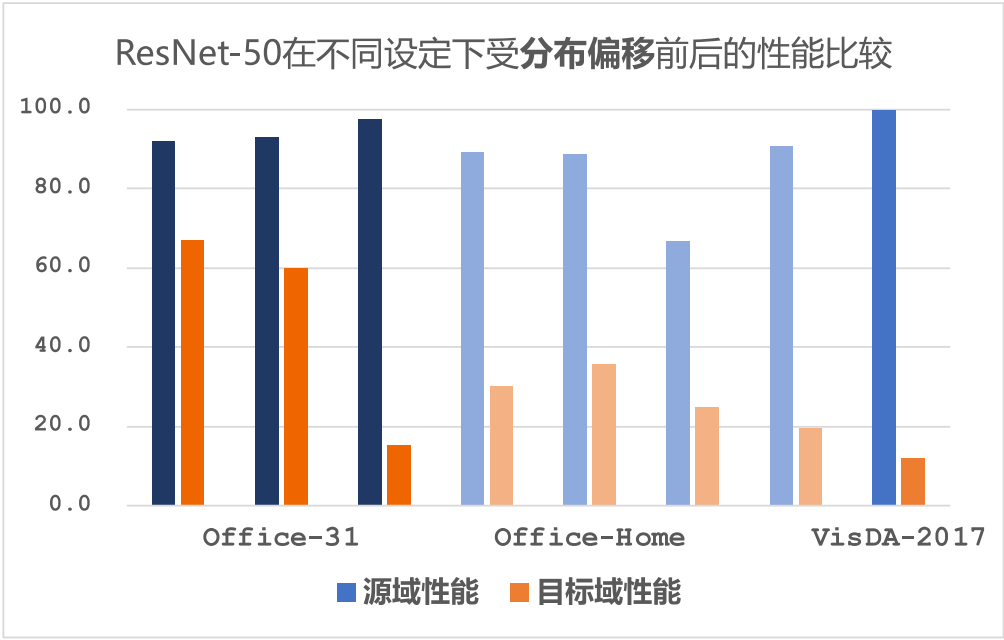
理论分析

独立同分布假设（IID Hypothesis）：**源域**P_s和**目标域**P_t（应用场景）独立同分布。在此情况下进行知识蒸馏，源域的知识可以很好地指导模型完成目标域的任务。

- 数据蒸馏的目标：

$$\min_{\theta_s} \mathbb{E}_{(x,y)\sim \boldsymbol{P}}[D_{KL}(T(x;\theta_t) \parallel S(x;\theta_s) + CE(S(x;\theta_s), y)].$$

- 多数场景下，源域分布和应用场景存在**分布偏移**（$P_t \neq P_s$），违反独立同分布假设。
  - 情况1：$\boldsymbol{P} \approx P_t$，对应无数据蒸馏方法（$P_t$ 由生成器拟合），蒸馏出的目标模型并不适用P_s。
  - 情况2：$\boldsymbol{P} \approx P_s$，源模型给出的知识不一定有效。



Zihao Tang, Zheqi Lv, Shengyu Zhang, Yifan Zhou, Xinyu Duan, Fei Wu, Kun Kuang: AuG-KD: Anchor-Based Mixup Generation for Out-of Domain Knowledge Distillation. ICLR 2024

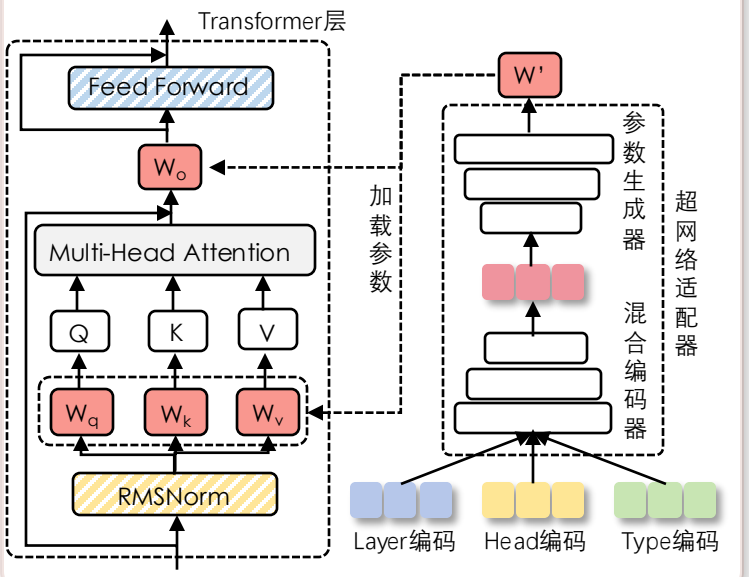
运行时大模型极限压缩技术 – HyperCAT

研究问题 如何在资源受限的边缘设备上对LLMs进行极限压缩，同时保持模型性能和表达能力？

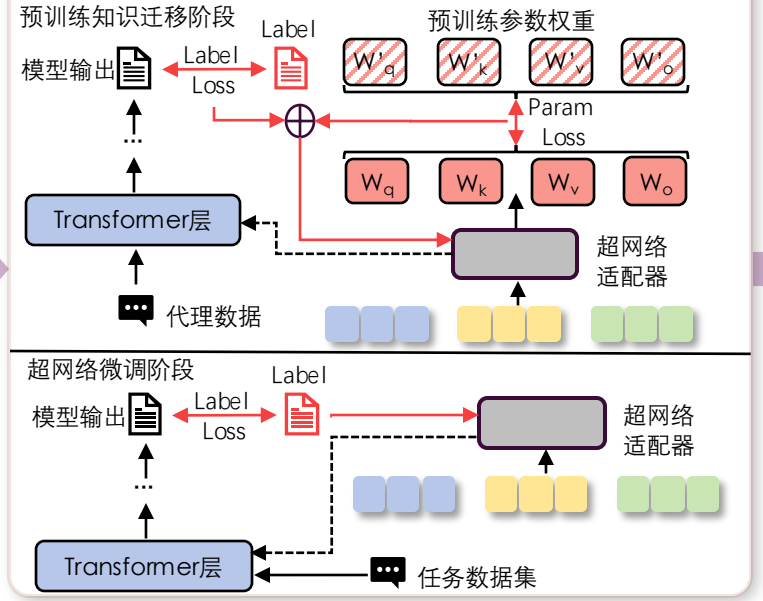
解决方案

- 引入轻量级超网络适配器，动态生成层特定的注意力模块参数，**运行时显存永远只需存储一层网络！**
- 通过预训练知识迁移与超网络微调结合的二阶段训练，确保生成参数保留模型的通用能力，又能适配下游任务。

超网络参数生成



二阶段训练调优



实验结果

Table 1: Performance of HyperCAT compared to traditional fine-tuning methods. The best results for each setting are highlighted in bold.

Methods	SST-2	RTE	BoolQ
LP	92.75	61.87	61.19
Partial FT	93.58	63.30	59.63
HyperCAT	**95.39**	**64.03**	**62.75**

Table 2: Hypernetwork capacity analysis through layer-Wise parameter generation. The best results for each setting are highlighted in bold.

Methods	Pre-KT	Hyper-FT
HyperCAT (No Pre-KT)	-	50.92
HyperCAT (all layers)	76.56	79.08
HyperCAT (last 14 layers)	92.86	92.75
HyperCAT (last 4 layers)	93.08	94.55
HyperCAT (last 3 layers)	93.90	94.81
HyperCAT (last 2 layers)	93.41	94.35
HyperCAT (last 1 layer)	94.02	95.39

成效 降低模型存储内存占用（40%），以少量可训练参数（3%）实现高效微调，实现优于传统方法的性能

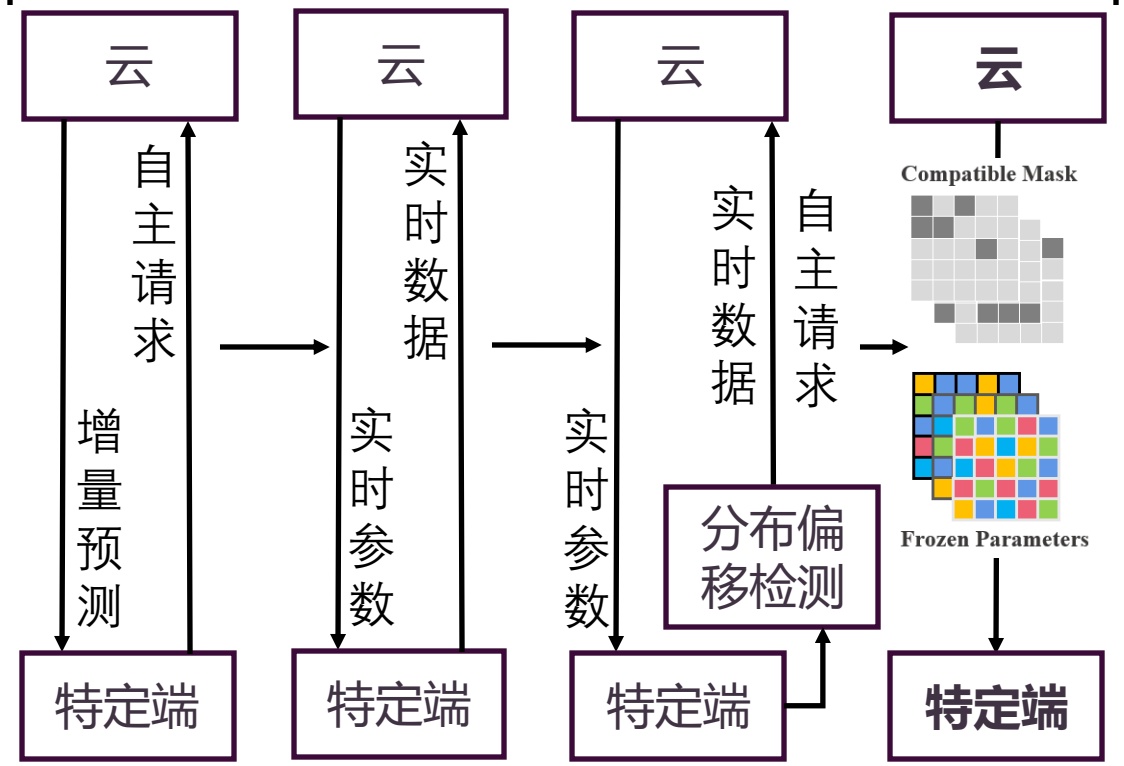
Kunxi Li, Zhouzhou Shen, Wenyan Fan, Shengyu Zhang. HyperCAT: Dynamic Hypernetwork-Driven Parameter Compression for Edge-Compatible LLM Fine-Tuning. CEFSW @ ICMR 2025

端云异构模型知识互迁与协同推断

解决端云大小模型在差异化**尺寸架构**和**优化目标**下的**协同推断**问题

云不直接执行任务本身，而是帮助端更好的执行既定任务

Cloud for Device (C4D)



AdaRequest
自主请求
KDD 22

DUET
实时适应
WWW 23

IntellectReq
实时自主适应
WWW 24

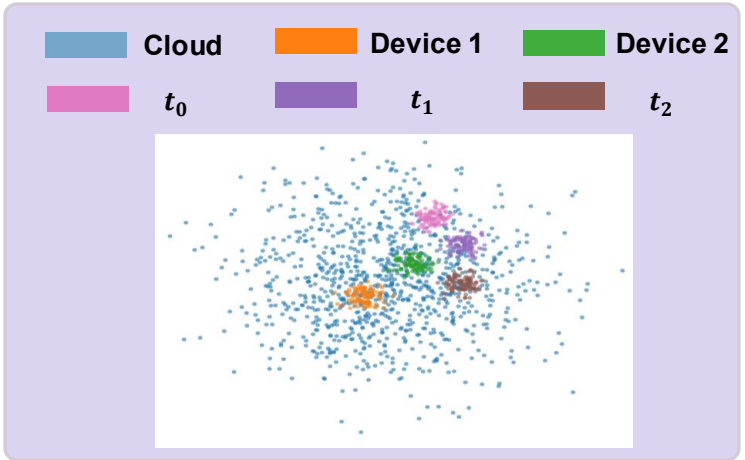
DIET/
Forward-OFA
高效定制
KDD 24/KDD 25

基于端云协同的高效端模型参数定制

研究背景 现有端侧部署方案采用云侧大规模预训练，通过模型压缩后传输至端侧进行部署。然而多阶段训练、稠密信息传输给端侧动态复杂环境下的**高响应、低成本**自适应带来了巨大挑战

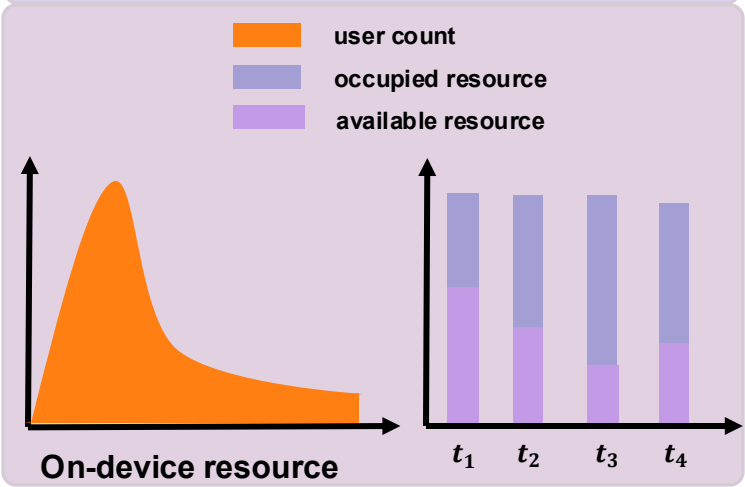
分布异质性

- **端云分布异质**：云侧全局数据分布体现平台整体共性与端侧特化分布存在偏移
- **端侧分布迁移**：端侧用户兴趣意图动态偏移，需要由云向端及时下发适配模型



资源异质性

- **端侧计算资源有限**：大量长尾用户移动设备算力有限，难以支撑本地训练微调
- **端云通信资源有限**：频繁下发稠密适配模型消耗大量通信带宽资源，降低响应



Fu K, Zhang S, Lv Z, et al. DIET: Customized Slimming for Incompatible Networks in Sequential Recommendation. KDD 2024 Research Track

基于端云协同的高效端模型参数定制

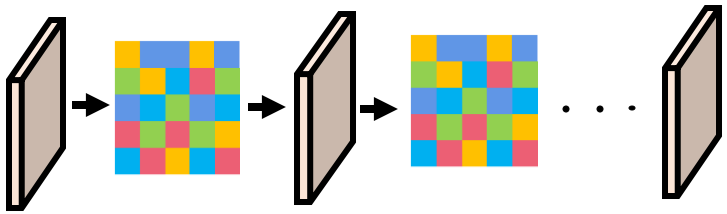
研究问题　研究基于端云协同的低通信开销、高响应速度端模型定制算法。

创新方法

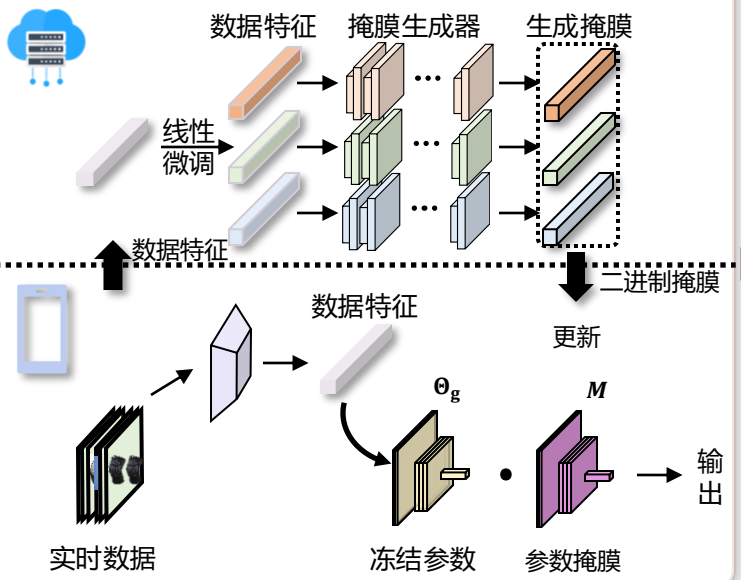
高效模型表示构建：基于神经网络彩票假说，将云向端训练压缩过程转化为传输适配子网二进制掩膜

高效适配子网搜索：云侧学习建立实时数据到端侧个性子网掩膜的映射，仅需前向推理即可高效响应

彩票假说理论

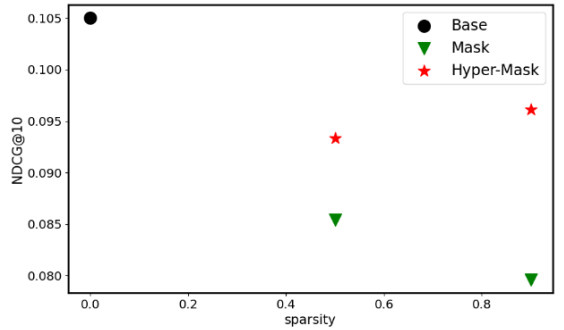


端云子网搜索



模型效率提升

优势 \ 方法	Base	Ours
低传输延迟	✗	✓
低存储成本	✗	✓
低推理时延	✗	✓



低时延低成本下得到相似的表现

Fu K, Zhang S, Lv Z, et al. DIET: Customized Slimming for Incompatible Networks in Sequential Recommendation. KDD 2024 Research Track

基于端云协同的高效端模型参数定制

应用验证

降低模型由云向端下发的传输开销至**原始大小的3%**
端侧模型能力提升的同时**推理速度提升5倍**

Model	Method	Dataset							
		Amazon-CD				MovieLens-100k			
		NDCG↑	Hit↑	Params↓	FLOPs↓	NDCG↑	Hit↑	Params↓	FLOPs↓
SASRec	Base	0.0386	0.0529	1.3107	0.2086	0.0517	0.1077	1.3107	0.2086
	DIET	**0.0425**	**0.0590**	**0.0410**	**0.1154**	**0.0635**	**0.1319**	**0.0410**	**0.0416**
	Improv. ↑	10.96%	11.53%	× 31.97	× 1.81	22.82%	22.47%	× 31.97	× 5.01
Caser	Base	0.0310	0.0424	0.4922	0.0586	0.0569	0.0719	0.4922	0.0586
	DIET	**0.0356**	**0.0488**	**0.0154**	**0.0294**	**0.0617**	**0.0771**	**0.0154**	**0.0488**
	Improv. ↑	14.84%	15.09%	× 31.96	× 1.99	10.42%	7.32%	× 31.96	× 1.76

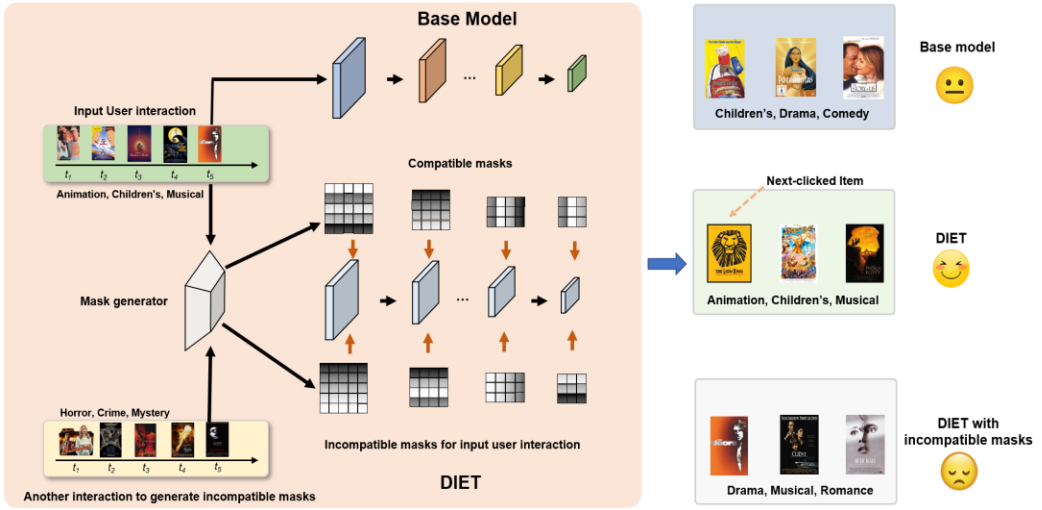
当前推荐系统存在的问题

通信开销大
云端分布差异大
端侧兴趣变化快
设备计算资源有限

瘦身子网模型压缩
端侧实时兴趣提取
适配子网生成传输
…

端侧个性子网搜索

共性-个性协同
大-小模型协同
…



突破了端云协同计算在**分布偏移、资源受限**设备上训练推理效率局限

参数定制 → **结构定制**

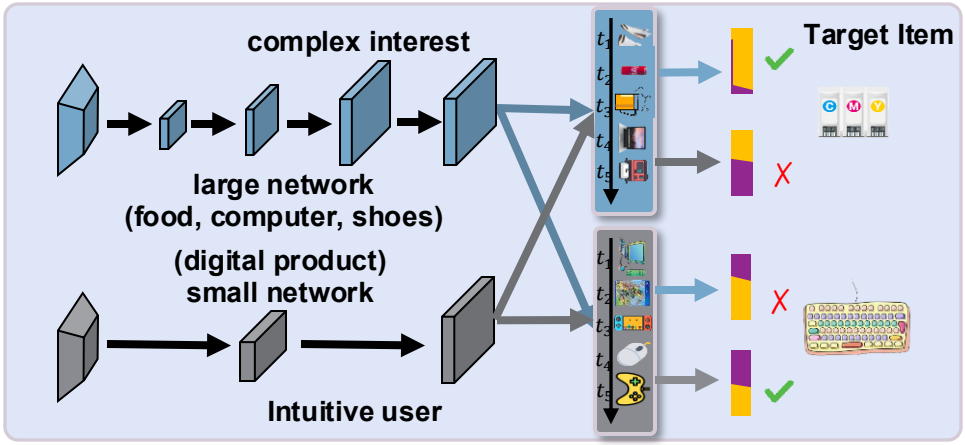
基于端云协同的高效端模型结构定制

研究背景

现有端侧部署方案主要关注端侧**模型参数**个性化，通过为每个设备端定制不同的参数来适应变化的兴趣。然而，这类方案所采用的**统一模型结构**给端侧差异化动态资源限制下的**高响应、低成本**带来了巨大挑战。

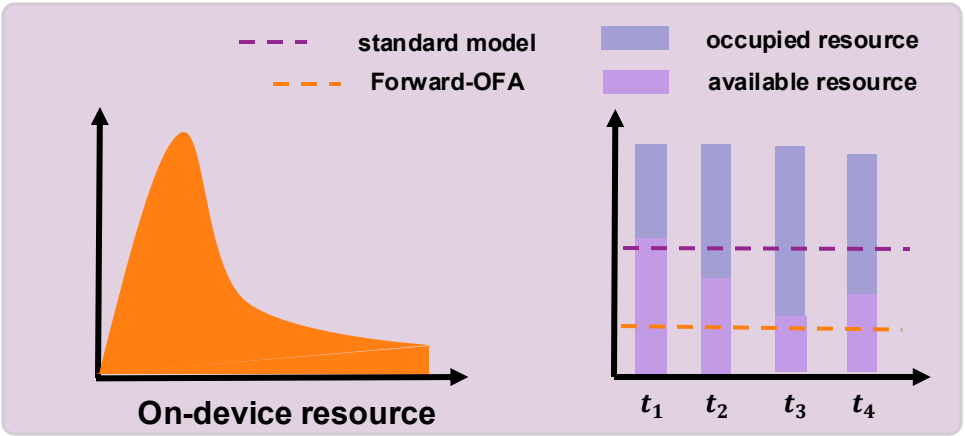
分布异质性

- **端云分布异质**：云侧全局数据分布体现平台整体共性与端侧特化分布存在偏移
- **用户结构偏好**：较小的模型足以用来建模直观用户兴趣，而更深层次的兴趣需要使用较大的模型来建模



资源异质性

- **端侧计算资源动态化**：移动端会同时运行其他进程，模型推理可利用的资源随着时间动态变化
- **端侧计算资源差异大**：不同设备之间的计算资源有较大差异，大部分端计算资源匮乏



Fu K, Lv Z, Zhang S, et al. Forward Once for All: Structural Parameterized Adaptation for Efficient Cloud-coordinated On-device Recommendation[C]//Proceedings of the 31th ACM SIGKDD Conference on Knowledge Discovery & Data Mining.

基于端云协同的高效端模型结构定制

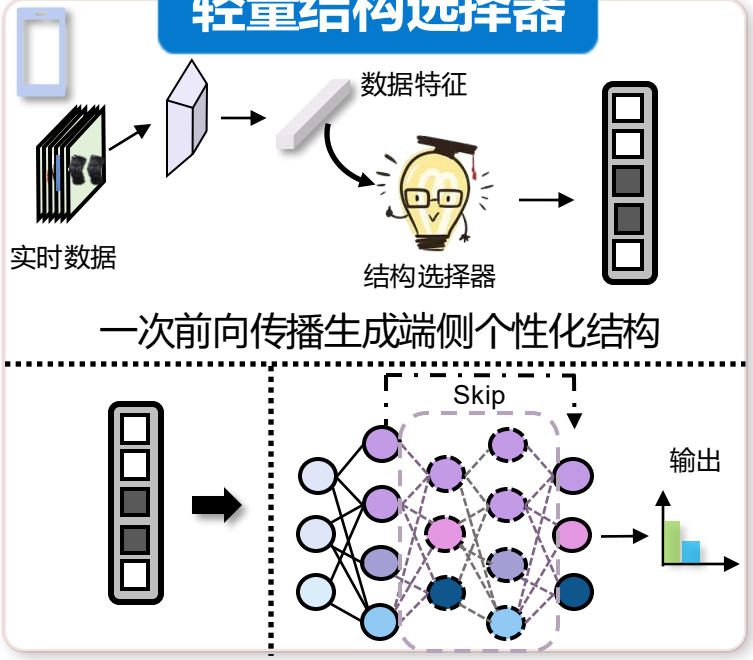
研究问题 研究基于端云协同的低端侧负载、高推理速度端轻量模型结构定制算法。

创新方法

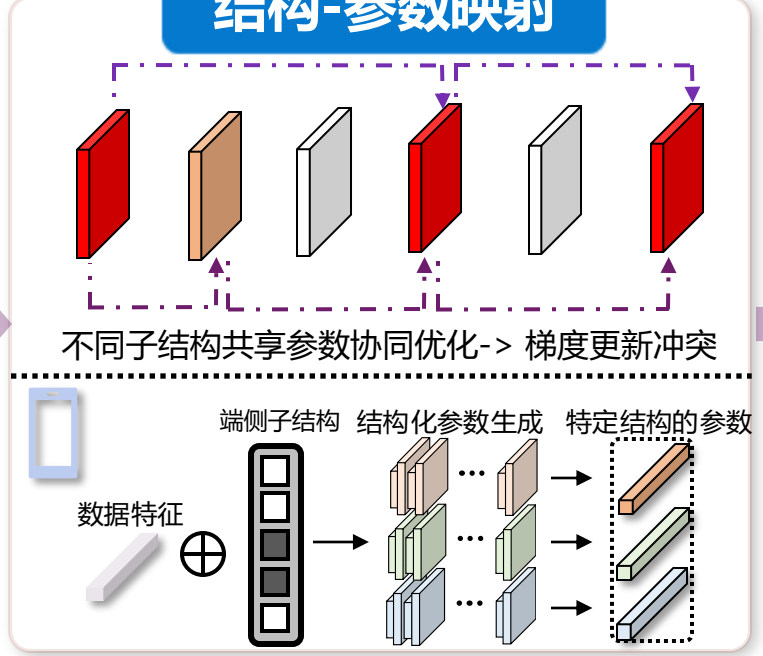
高效推理路径构建：根据端侧实时数据，在一次前向传播中推理得到适配端侧数据分布的模型结构

结构匹配参数映射：云侧构建从模型结构到特定参数的映射，避免不同子结构共用参数带来的梯度冲突

轻量结构选择器



结构-参数映射



模型效率提升

方法 / 优势	On-device RecSys	Pruning RecSys	Ours
个性化	✓	✗	✓
低传输时延	✓	✓	✓
快速推理	✗	✓	✓
快速适应	✗	✓	✓

低推理成本的同时得到更好的表现

基于端云协同的高效端模型结构定制

应用验证

端侧模型能力提升的同时**推理速度提升**

Model	Method	Dataset							
		MovieLens-1M				Amazon-Food			
		NDCG ↑	Hit ↑	FLOPs ↓	Param ↓	NDCG ↑	Hit ↑	FLOPs ↓	Param↓
SASRec	DeviceRec	0.0969	0.1816	0.6244	3.9936	0.0526	0.0620	0.6244	3.9936
	Finetune	0.0939	0.1793	0.6244	3.9936	0.0523	0.0623	0.6244	3.9936
	Forward-OFA	**0.1182**	**0.2081**	**0.5699**	**3.6384**	**0.0590**	**0.0713**	**0.5667**	**3.6160**
	Improv.↑	21.97%	14.59%	× 1.10	× 1.10	12.29%	14.90%	× 1.10	× 1.10
NextItNet	DeviceRec	0.0975	0.1846	1.5053	9.5846	0.0401	0.0467	1.5053	9.5846
	finetune	0.0879	0.1715	1.5053	9.5846	0.0399	0.0467	1.5053	9.5846
	Forward-OFA	**0.1226**	**0.2140**	**0.6167**	**3.8656**	**0.0523**	**0.0613**	**0.3126**	**1.9104**
	Improv.↑	25.74%	15.96%	×2.44	×2.48	30.24%	30.99%	×4.82	×5.02

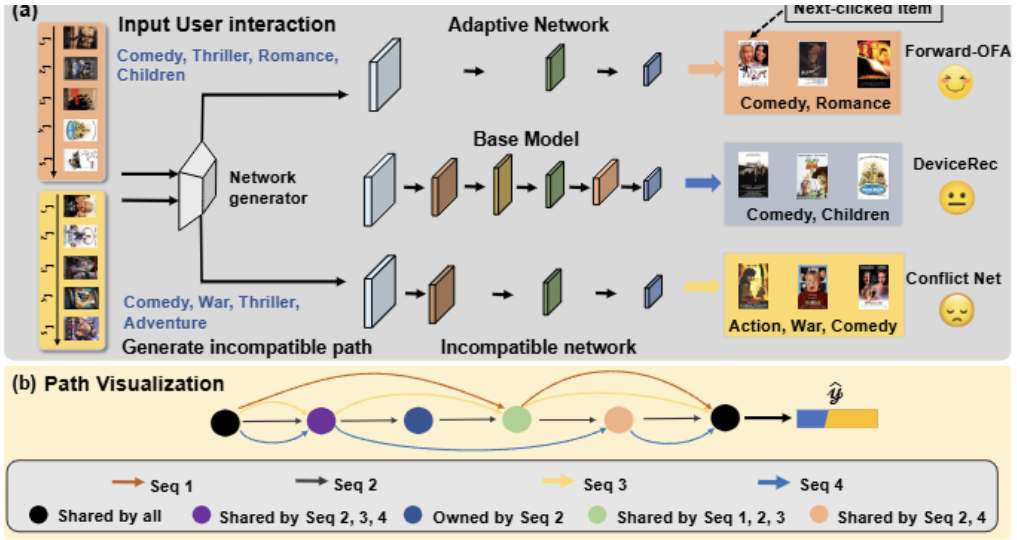
端侧结构个性定制
结构参数协同进化
端侧实时兴趣提取
…

端侧子结构定制

共性-个性协同
大-小模型协同
…

当前推荐系统存在的问题

动态设备可用资源
差异化端计算资源
云端分布差异大
端侧兴趣变化快



突破了端云协同计算在**分布偏移、资源受限**设备上推理效率局限

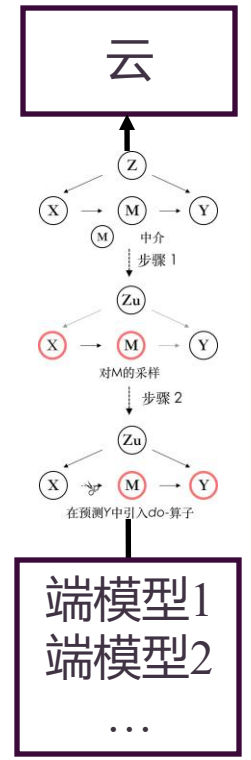
端云异构模型知识互迁与协同推断

Cloud to Device (C2D)

Cloud for Device (C4D)

Device to Cloud (D2C)

面对**不同端数据偏差和训练偏差下**，解决个性化端模型向去偏云汇聚学习难题。

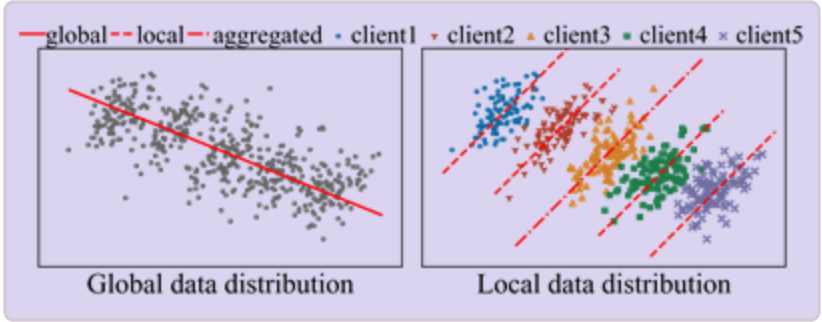


FedCFA/CausalD
因果去偏汇聚
AAAI 25, TKDE 23

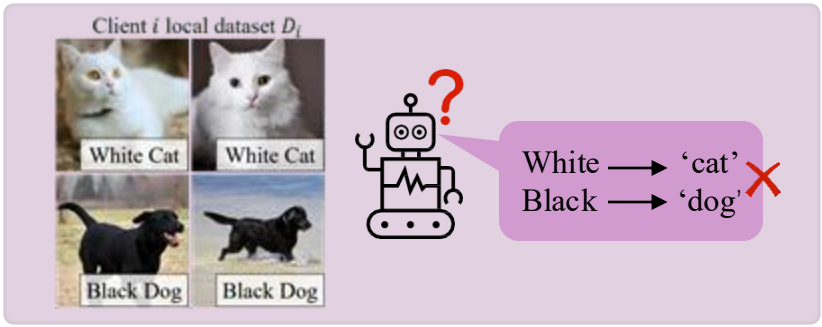
利用端侧反事实表征学习实现端向云去偏汇聚

研究背景 数据分布异质性导致的"局部观察到的趋势在全局数据中消失或反转"的辛普森悖论，使得云侧汇聚模型无法准确反映整体数据分布，给端向云去偏汇聚带来了巨大挑战

分布异质

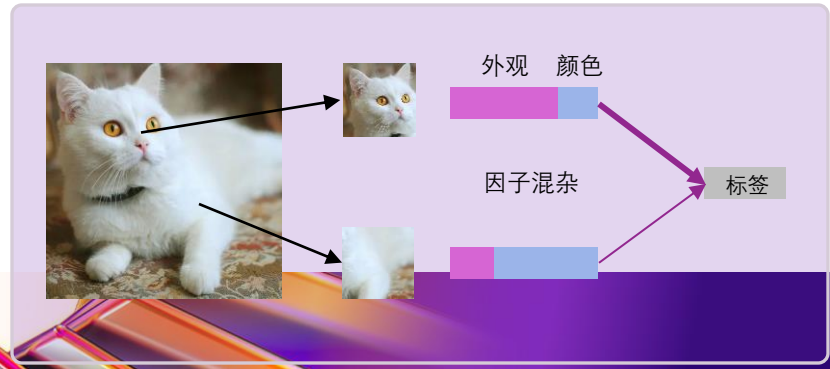


- **端云分布异质**：云侧全局数据分布体现平台整体共性与端侧特化分布存在偏移
- **端云有偏汇聚**：有偏数据导致端侧偏见，相似偏见端侧模型导致云侧有偏汇聚



因子混杂

- **虚假相关**：端侧数据局部且有限，存在虚假的因子-标签关联，忽视真实因果关系
- **因子耦合**：因子之间存在复杂的相互依赖关系，难以有效解耦出独立的因果关系



Jiang Z, Xu J, **Zhang S**, et al. FedCFA: Alleviating Simpson's Paradox in Model Aggregation with Counterfactual Federated Learning. AAAI 2025

利用端侧反事实表征学习实现端向云去偏汇聚

研究问题

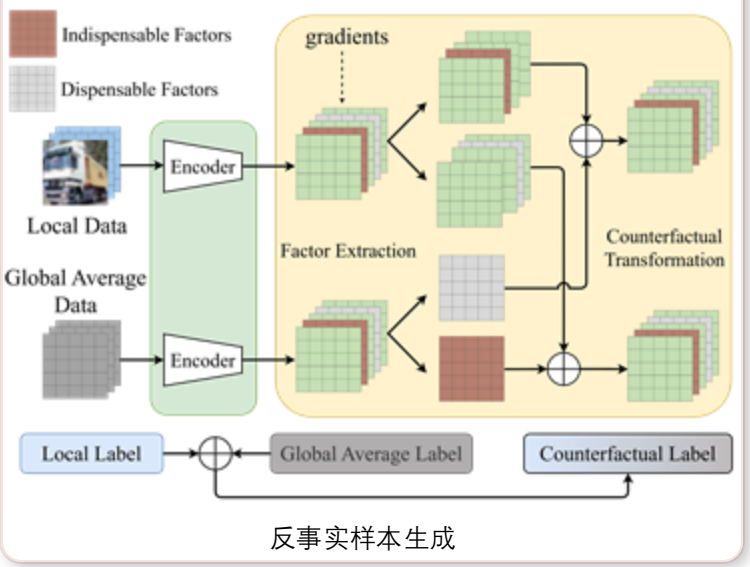
利用端侧反事实表征学习解决云侧模型联邦汇聚中“辛普森悖论”难题。

创新方法

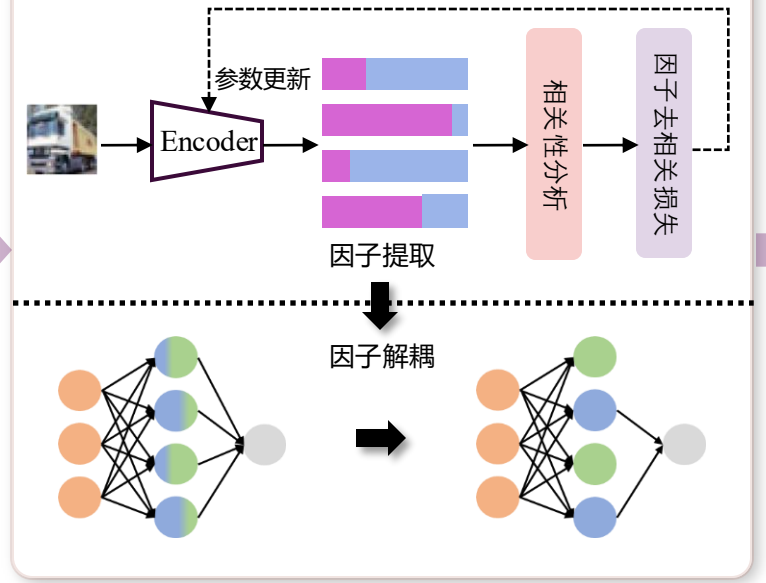
反事实表征学习：利用全局平均数据信息在端侧生成反事实样本，实现端侧模型去偏训练

因子去相关模块：基于相关性分析设计因子去相关模块对因子解耦，提高反事实样本的质量

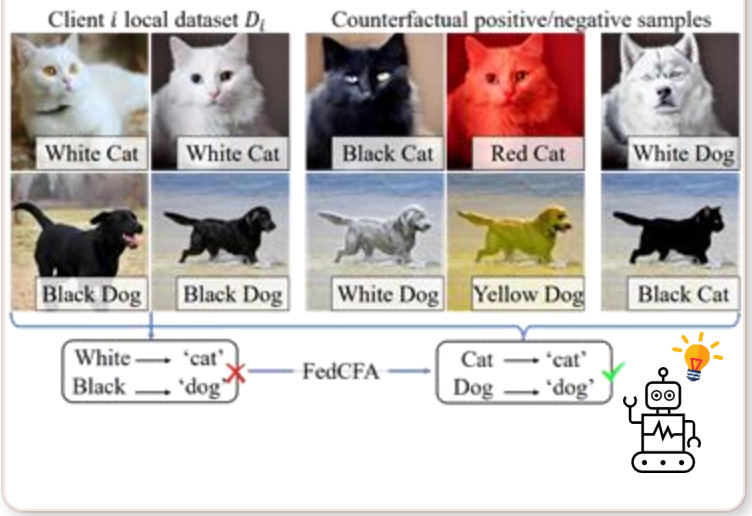
反事实表征学习



因子去相关模块



缓解辛普森悖论



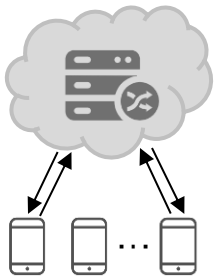
Jiang Z, Xu J, **Zhang S**, et al. FedCFA: Alleviating Simpson's Paradox in Model Aggregation with Counterfactual Federated Learning. AAAI 2025

利用端侧反事实表征学习实现端向云去偏汇聚

实验验证

相比于主流联邦学习的最佳方法，云侧模型**精度最高可提升7.75%**
云侧模型去偏汇聚的同时**收敛速度提升2倍**

当前端云协同存在的问题



数据高度异质性
云端分布差异大
云侧模型收敛慢

反事实样本生成
因子去相关约束
混杂因子解耦合
…

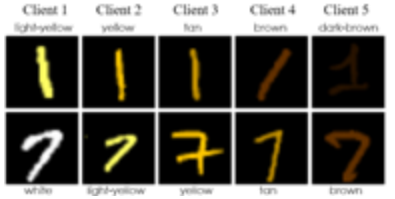
端侧反事实表征学习

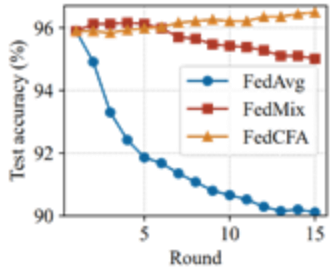
端-云模型协同
…

	Method	$Dir_{60}(0.2)$	$Dir_{60}(0.6)$	IID_{60}	$Dir_{100}(0.2)$	$Dir_{100}(0.6)$	IID_{100}
CIFAR100	FedAvg	40.70±0.24	42.85±0.26	44.37±0.40	38.17±0.32	40.19±0.26	42.19±0.52
	FedProx	40.39±0.23	42.51±0.34	44.21±0.64	38.27±0.46	39.90±0.42	42.24±0.98
	SCAFFOLD	29.36±0.39	33.30±0.48	37.96±0.26	23.25±0.54	29.98±0.19	32.77±0.25
	FedPRV	38.35±1.11	42.91±0.49	45.91±0.05	30.65±0.74	36.58±0.14	39.96±0.30
	q-FedAvg	40.34±0.60	42.61±0.63	44.43±0.28	38.15±0.48	40.20±0.10	42.04±0.65
	FedMix	42.51±0.28	44.16±0.26	45.65±0.31	39.78±0.07	41.43±0.84	43.63±0.64
	FedCFA	**46.96±1.04**	**49.32±0.20**	**48.31±0.53**	**46.71±0.59**	**49.18±0.75**	**47.86±1.22**
CIFAR10	FedAvg	65.88±0.32	73.95±0.16	75.43±0.51	62.87±0.12	70.99±0.70	72.82±0.35
	FedProx	72.23±0.44	77.68±0.03	76.93±0.28	70.36±0.75	75.47±0.53	73.36±0.47
	SCAFFOLD	33.05±5.57	54.58±3.95	75.96±0.57	34.69±2.16	56.17±1.91	71.84±0.77
	FedPRV	59.42±2.26	71.52±0.21	77.42±0.04	55.43±1.74	67.11±0.76	76.16±0.37
	q-FedAvg	71.71±1.05	77.96±0.19	76.92±0.09	70.04±1.55	75.47±0.52	73.68±0.33
	FedMix	74.61±0.74	78.64±0.53	77.90±0.17	73.91±0.79	77.11±0.31	73.93±0.06
	FedCFA	**75.89±1.00**	**82.43±0.08**	**83.36±0.51**	**75.76±0.15**	**81.73±0.12**	**81.68±0.89**

Method	Tiny-ImageNet		FEMNIST	Sent140
	$Dir_{60}(0.2)$	$Dir_{60}(0.6)$	Non-IID	Non-IID
FedAvg	27.39±0.13	30.90±0.29	81.31±0.94	68.10±0.48
FedProx	27.34±0.05	30.78±0.16	81.63±0.08	68.10±0.44
q-FedAvg	26.89±0.07	30.70±0.13	81.90±0.58	68.04±0.71
FedMix	28.01±0.19	32.43±0.13	82.31±0.21	67.97±0.39
FedCFA	**30.70±0.68**	**32.86±0.77**	**83.19±0.54**	**69.26±0.37**

Method	$Dir_{60}(0.2)$	IID_{60}	$Dir_{100}(0.2)$	IID_{100}
Target	75.5	78.4	73.6	75.6
FedAvg	(>,66.22)	(>,74.99)	(>,62.95)	(>,72.42)
FedProx	(>,74.03)	(>,77.16)	(>,72.37)	(>,73.46)
SCAFFOLD	(>,32.77)	(693,78.46)	(>,38.34)	(800,75.68)
q-FedAvg	(>,72.23)	(>,77.04)	(>,72.03)	(>,73.56)
FedMix	(610,75.69)	(>,77.79)	(>,72.25)	(>,74.00)
FedCFA	**(375,75.58)**	**(453,78.41)**	**(427,73.61)**	**(408,75.67)**



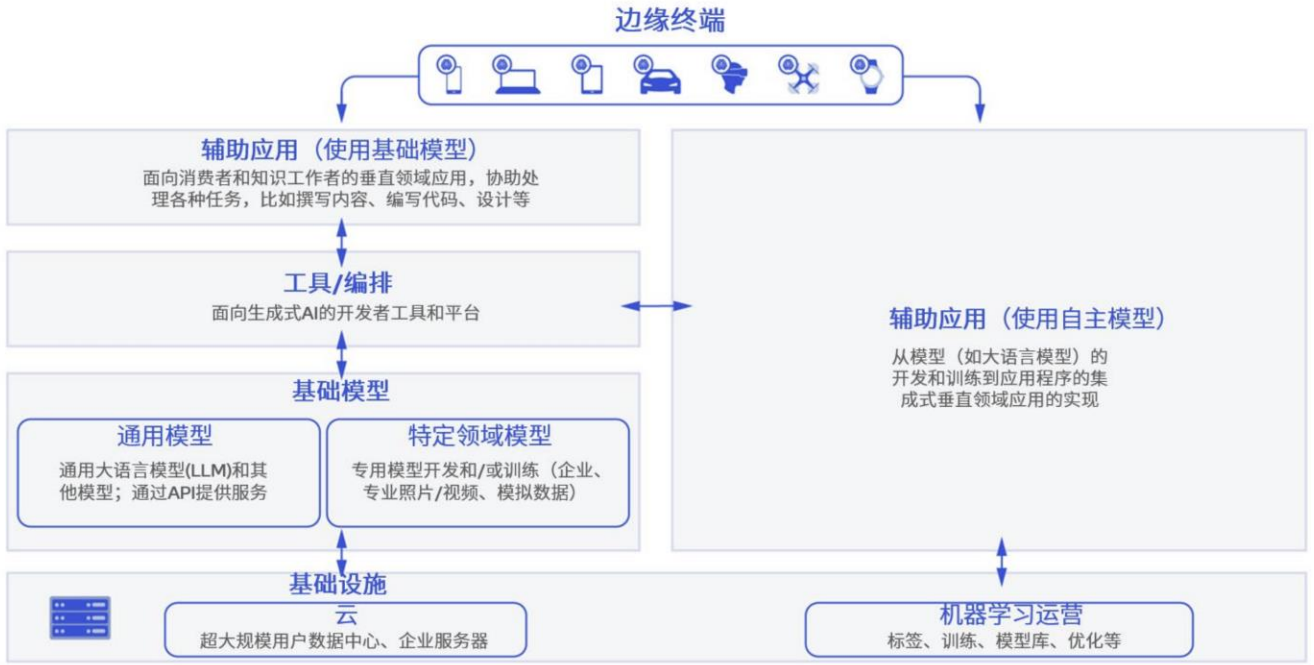


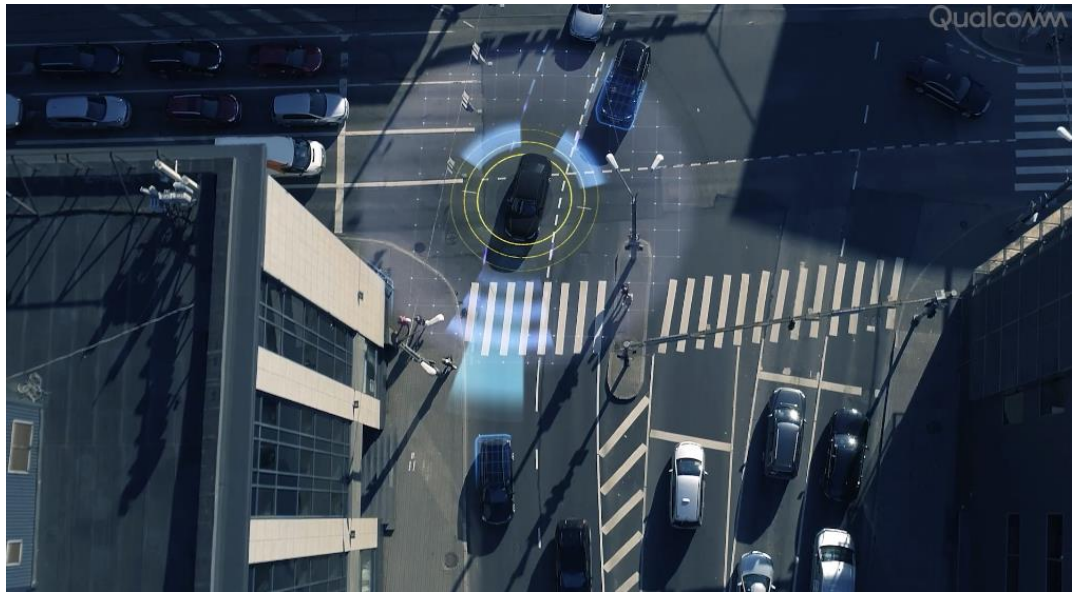
突破了端云协同计算在**分布偏移、数据异质**场景中
模型汇聚效率局限

PART 04

应用分析

高通：生成式端云混合智能



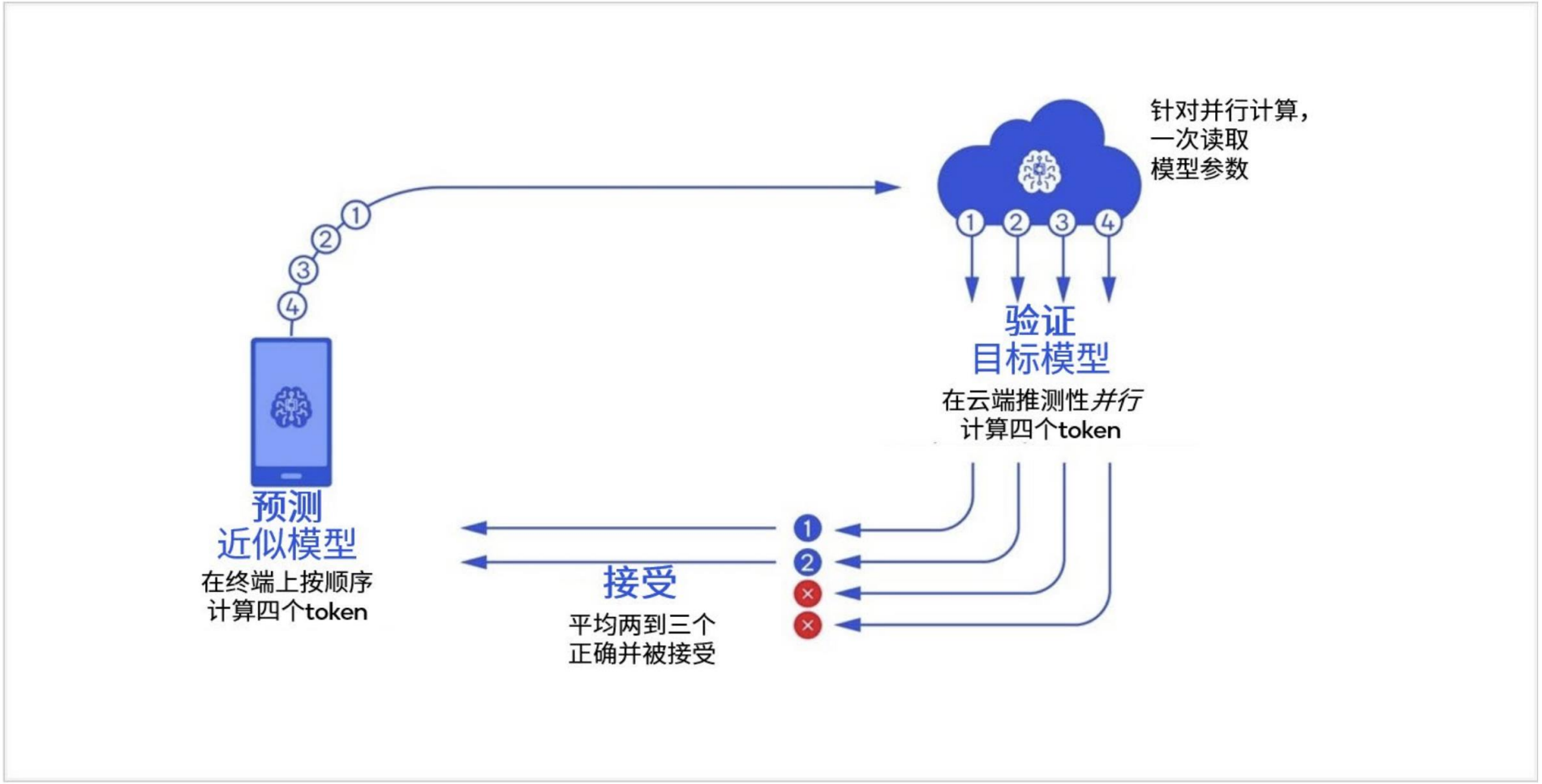


- 混合AI指终端和云端协同工作，在适当的场景和时间下分配AI计算的工作负载，以提供更好的体验，并高效利用资源。在一些场景下，计算将主要以终端为中心，在必要时向云端分流任务。而在以云为中心的场景下，终端将根据自身能力，在可能的情况下从云端分担一些AI工作负载。

端云协同智能

-- 高通《终端侧AI 和混合AI 开启生成式AI 的未来》

端云协同智能

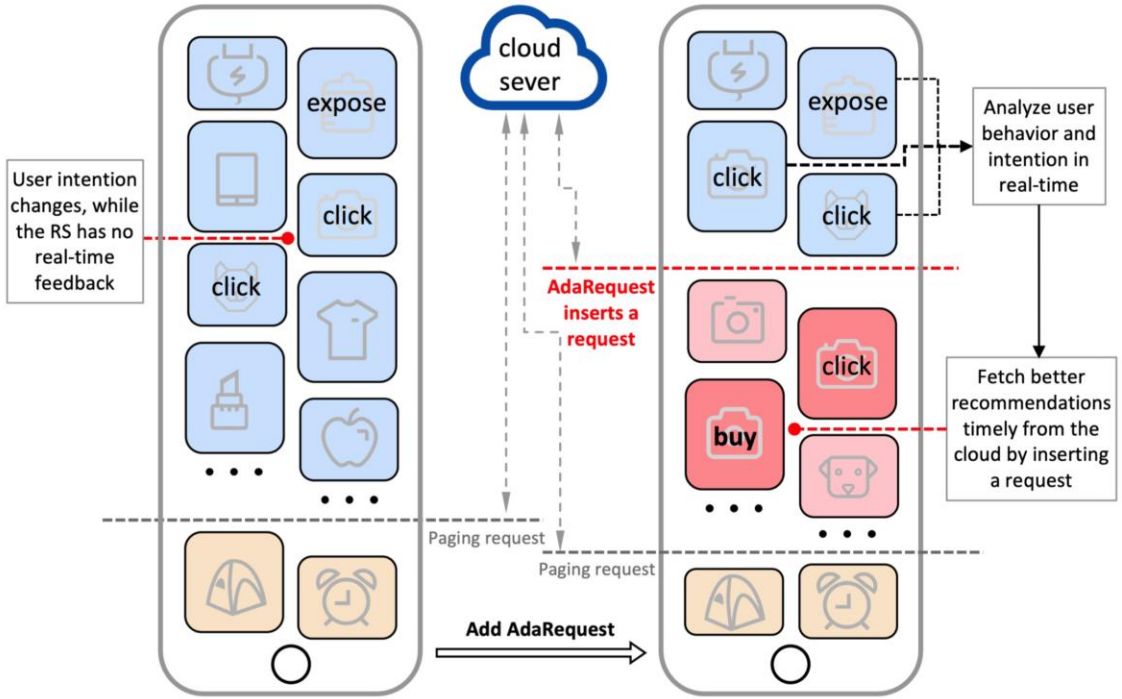


-- 高通《终端侧AI 和混合AI 开启生成式AI 的未来》

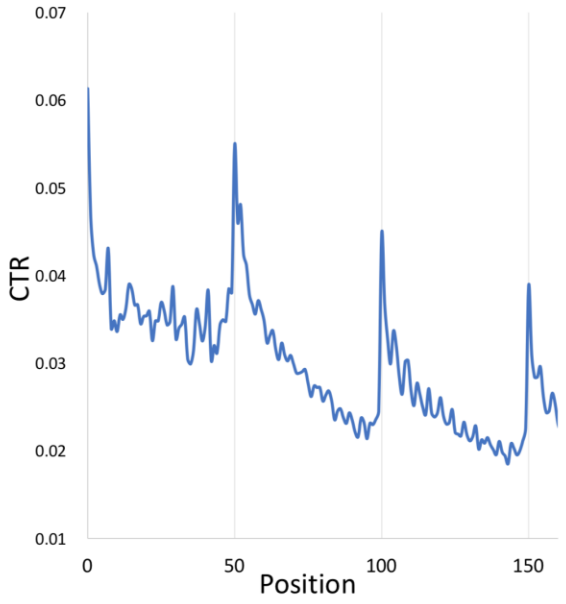
端云大-小模型协同推断算法

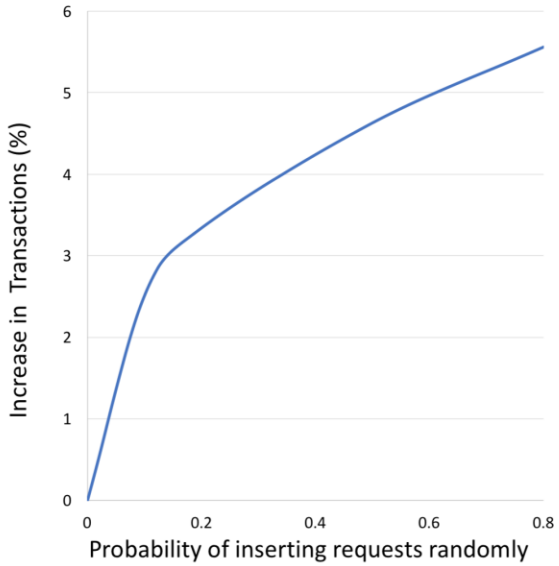
- 动态变化的端环境导致资源有限情况下云模型的延迟响应，导致端侧服务与端侧环境的不匹配，损害用户的服务体验

手机淘宝商品推荐系统



用户点击率在云模型响应后陡升

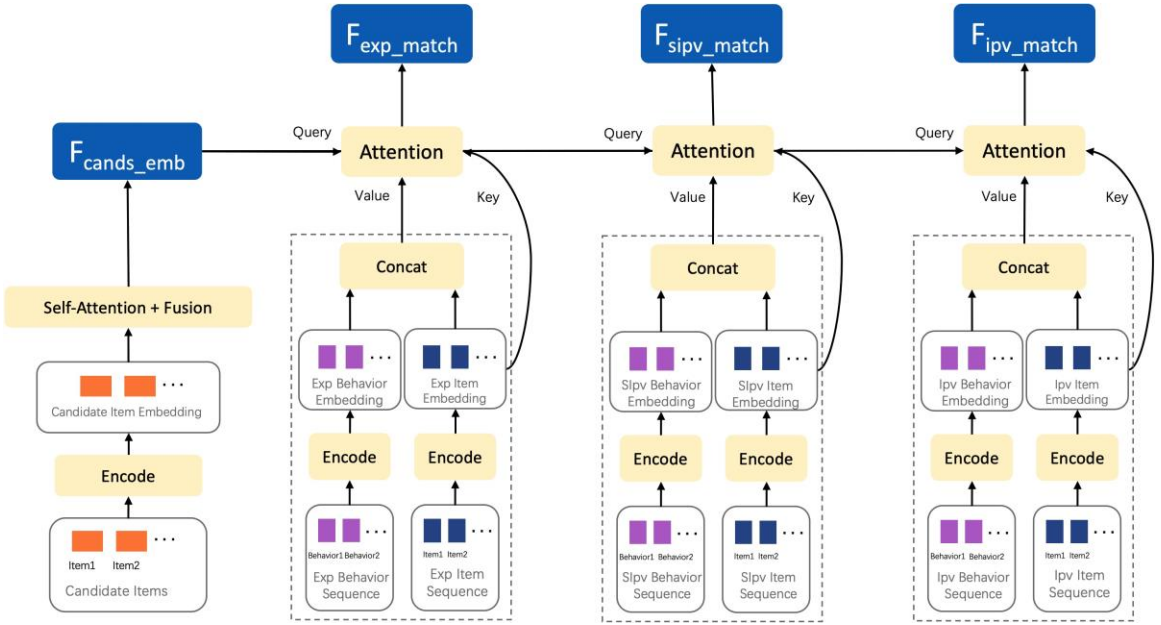


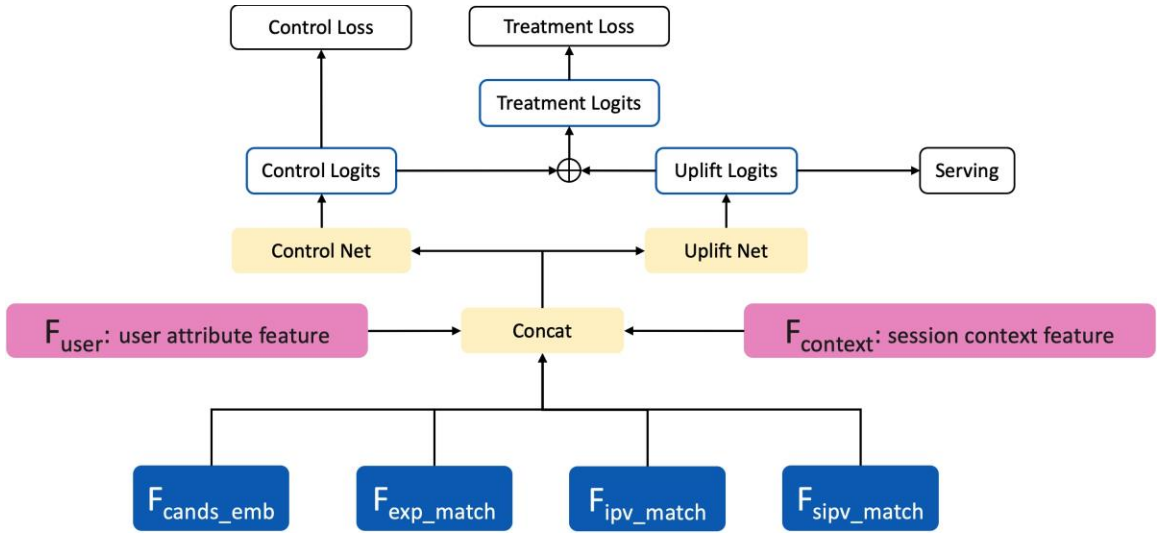


Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang*, Ziwen Jiang, Qingwen Liu, Xiaoyi Zeng, Tat-Seng Chua, Fei Wu. Intelligent Request Strategy Design in Recommender System, KDD 2022

端云大-小模型协同推断算法

- 端设备部署小模型实时检测端环境变化（用户兴趣意图变化）
- 通过因果潜在结果模型预估请求大模型响应价值
- 动态规划对云侧大模型的请求,最大化资源有限时的线上收益。





Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang*, Ziwen Jiang, Qingwen Liu, Xiaoyi Zeng, Tat-Seng Chua, Fei Wu. Intelligent Request Strategy Design in Recommender System, KDD 2022

端云大-小模型协同推断算法

当前推荐系统存在的问题

通信开销大
隐私破坏风险
隐时反馈噪声多
无法实时感知用户

因果结构学习机制
因果潜在结构框架
不确定性预估方法
…

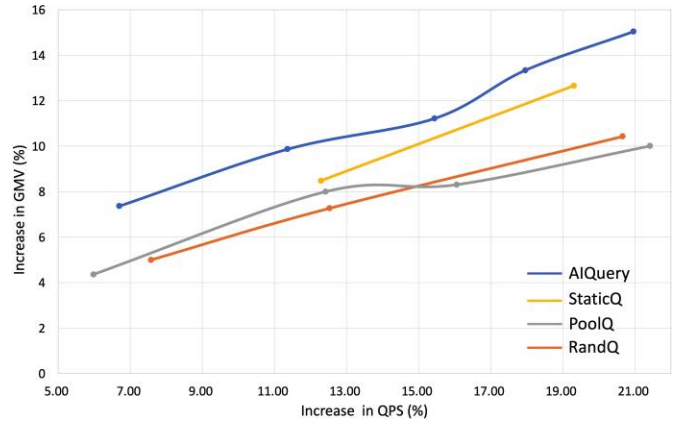
因果+端云协同

共性-个性协同
大-小模型协同
隐私-效率协同
…

直接经济效益
(购买率)

PR in N	NoQ	RandQ	PoolQ	StaticQ	AIQuery
10	0.889	1.174	1.177	1.130	**2.289**
20	1.660	2.197	2.164	2.045	**4.173**

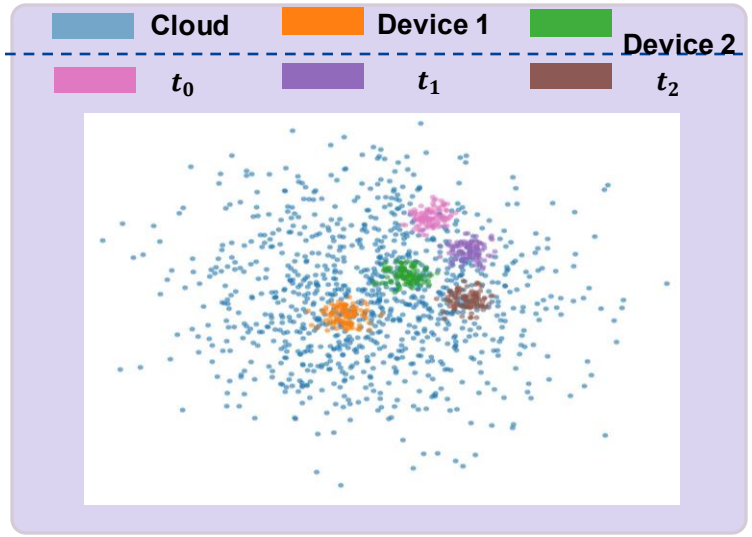
平台经济效益
(商品交易总值)



Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang*, Ziwen Jiang, Qingwen Liu, Xiaoyi Zeng, Tat-Seng Chua, Fei Wu. Intelligent Request Strategy Design in Recommender System, KDD 2022

云上大语言模型和端上小推荐模型的端云协同推荐

研究背景 现有大模型应用于推荐系统可以集成多任务且性能强大，然而面对端上复杂多变的用户行为，大的模型尺寸和高的运算成本为端侧推荐带来了**难部署**和**高延迟**的挑战。

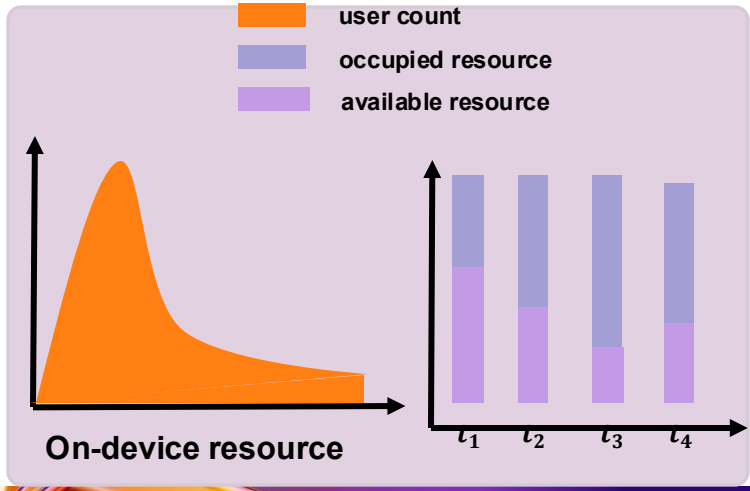


数据异质性

- **端云分布异质**：云侧全局数据分布体现平台整体共性与端侧特化分布存在偏移
- **端侧分布迁移**：端侧用户兴趣意图动态偏移，需要由云向端及时下发适配候选列表

资源异质性

- **端侧存储与计算资源**：用户移动设备存储和算力有限，难以部署和运算大模型
- **云侧存储与计算资源**：云上的存储资源可以部署大模型，但大模型频繁训练和推理依然会消耗大量资源
- **端云通信资源有限**：频繁下发候选列表消耗大量通信带宽资源，降低响应速率



Lv Z, Zhan T, Wang W, et al. Collaboration of Large Language Models and Small Recommendation Models for Device-Cloud Recommendation[C]// In SIGKDD Conference on Knowledge Discovery and Data Mining (***KDD***), 2025.

云上大语言模型和端上小推荐模型的端云协同推荐

研究问题　研究基于端云协同的低通信开销、高响应速度端模型定制算法。

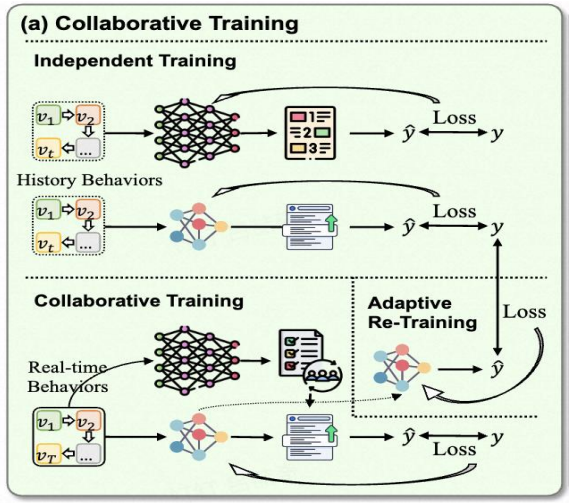
创新方法

协同训练：将云上大模型和端上小模型针对各自任务场景做针对性协作训练，提升场景适应性

协同推理：将云上大模型和端上小模型的输出结果融合，集成强泛化能力和强实时性的优势

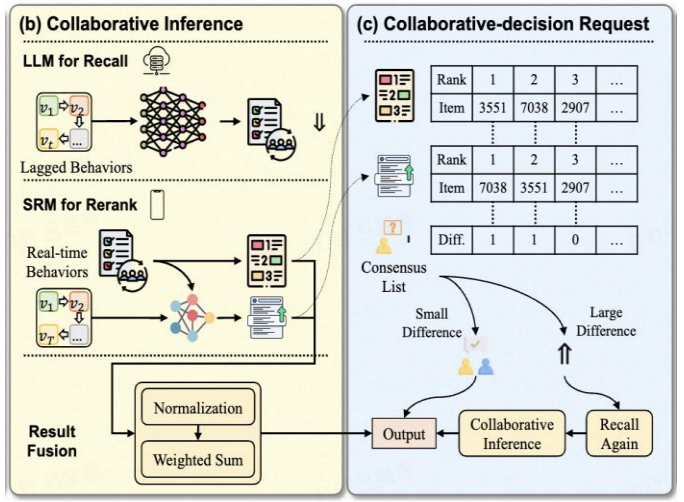
智能请求：对云上大模型和端上小模型的输出结果做不一致性检测，不一致性高的样本重新调用大模型

协同训练



大小协同训练，使小模型能针对大模型的候选列表有更强的排序能力

协同推理与请求



大小协同推理与请求，融合大小模型推理结果并决策何时调用云上大模型

模型效率提升

Dataset	Model	NDCG@5	NDCG@10	NDCG@20
Beauty	DIN (RT)	0.0079	0.0106	0.0135
	GRU4Rec (RT)	0.0081	0.0105	0.0129
	SASRec (RT)	0.0066	0.0096	0.0127
	P5 (RT)	0.0227	0.0257	0.0289
	DIN (NRT)	0.0017	0.0026	0.0042
	GRU4Rec (NRT)	0.0017	0.0023	0.0031
	SASRec (NRT)	0.0014	0.0021	0.0032
	P5 (NRT)	0.0087	0.0108	0.0136
	Ours (P5+SASRec)	**0.0094**	**0.0126**	**0.0154**
	Improve	8.41%	16.16%	13.45%
Toys	DIN (RT)	0.0046	0.0063	0.0081
	GRU4Rec (RT)	0.0050	0.0073	0.0098
	SASRec (RT)	0.0052	0.0073	0.0101
	P5 (RT)	0.0187	0.0204	0.0221
	DIN (NRT)	0.0007	0.0010	0.0017
	GRU4Rec (NRT)	0.0009	0.0015	0.0020
	SASRec (NRT)	0.0015	0.0020	0.0026
	P5 (NRT)	0.0066	0.0078	0.0090
	Ours (P5+SASRec)	**0.0066**	**0.0084**	**0.0096**
	Improve	0.46%	7.33%	6.76%

大幅补偿LLM无法获取实时数据下的推荐性能

Lv Z, Zhan T, Wang W, et al. Collaboration of Large Language Models and Small Recommendation Models for Device-Cloud Recommendation ***KDD* 2025**.

云上大语言模型和端上小推荐模型的端云协同推荐

应用验证

降低模型端云通信的传输开销至**原始大小的33%**
端侧模型**性能提升16.18%**

当前大模型应用于推荐系统存在的问题

通信开销大
端侧兴趣变化快
设备计算资源有限

端云大小模型
协同推荐性能对比

端云大小模型协同推荐

适用于多种大模型和小模型

Dataset	Model	Metric								
		NDCG@5	NDCG@10	NDCG@20	HR@5	HR@10	HR@20	Precision@5	Precision@10	Precision@20
Beauty	DIN (RT)	0.0079	0.0106	0.0135	0.0131	0.0226	0.0311	0.0026	0.0022	0.0018
	GRU4Rec (RT)	0.0081	0.0105	0.0129	0.0136	0.0212	0.0308	0.0026	0.0022	0.0017
	SASRec (RT)	0.0066	0.0096	0.0127	0.0111	0.0207	0.0310	0.0022	0.0021	0.0018
	P5 (RT)	0.0227	0.0257	0.0289	0.0305	0.0400	0.0525	0.0061	0.0040	0.0026
	DIN (NRT)	0.0017	0.0026	0.0042	0.0021	0.0039	0.0083	0.0014	0.0014	0.0015
	GRU4Rec (NRT)	0.0017	0.0023	0.0031	0.0020	0.0033	0.0054	0.0014	0.0012	0.0010
	SASRec (NRT)	0.0014	0.0021	0.0032	0.0018	0.0033	0.0060	0.0013	0.0011	0.0011
	P5 (NRT)	0.0087	0.0108	0.0136	0.0132	0.0200	0.0310	0.0027	0.0020	0.0016
	Ours (P5+SASRec)	**0.0094**	**0.0126**	**0.0154**	**0.0150**	**0.0248**	**0.0361**	**0.0030**	**0.0025**	**0.0018**
	Improve	8.41%	16.16%	13.45%	13.14%	23.91%	16.59%	13.21%	24.00%	16.77%

Dataset	LLM	Metric								
		NDCG@5	NDCG@10	NDCG@20	HR@5	HR@10	HR@20	Precision@5	Precision@10	Precision@20
Beauty	POD (NRT)	0.0079	0.0111	0.0134	0.0139	0.0239	0.0327	0.0028	0.0024	0.0016
	Ours(POD+SASRec)	**0.0091**	**0.0125**	**0.0150**	**0.0150**	**0.0256**	**0.0356**	**0.0030**	**0.0026**	**0.0018**
	Improv	15.59%	12.41%	12.27%	8.37%	7.29%	8.62%	8.30%	7.11%	8.54%

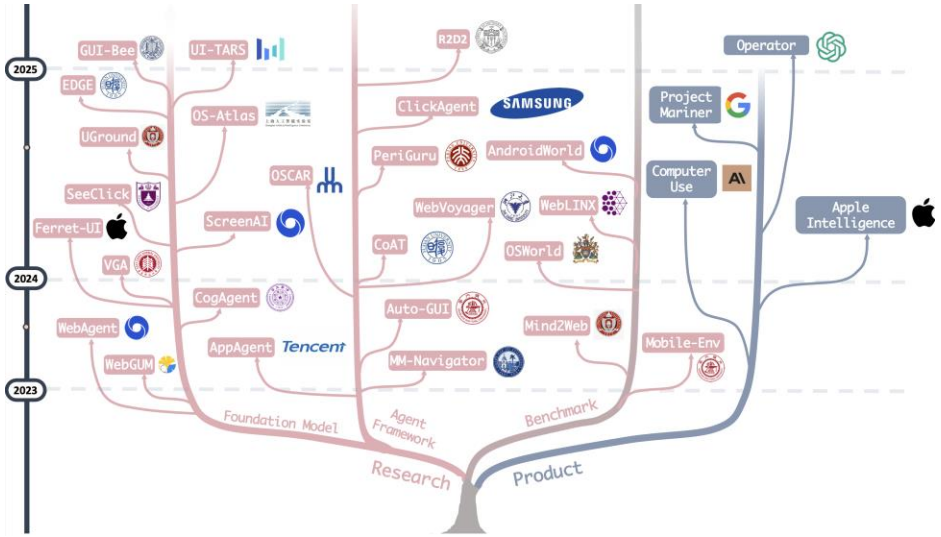
Dataset	SRM	Metric								
		NDCG@5	NDCG@10	NDCG@20	HR@5	HR@10	HR@20	Precision@5	Precision@10	Precision@20
Beauty	DIN	0.0091	0.0118	0.0147	0.0145	0.0230	0.0343	0.0029	0.0023	0.0017
	GRU4Rec	0.0092	0.0124	0.0151	0.0146	0.0247	0.0353	0.0029	0.0024	0.0017
	SASRec	**0.0094**	**0.0126**	**0.0154**	**0.0150**	**0.0248**	**0.0361**	**0.0030**	**0.0025**	**0.0018**

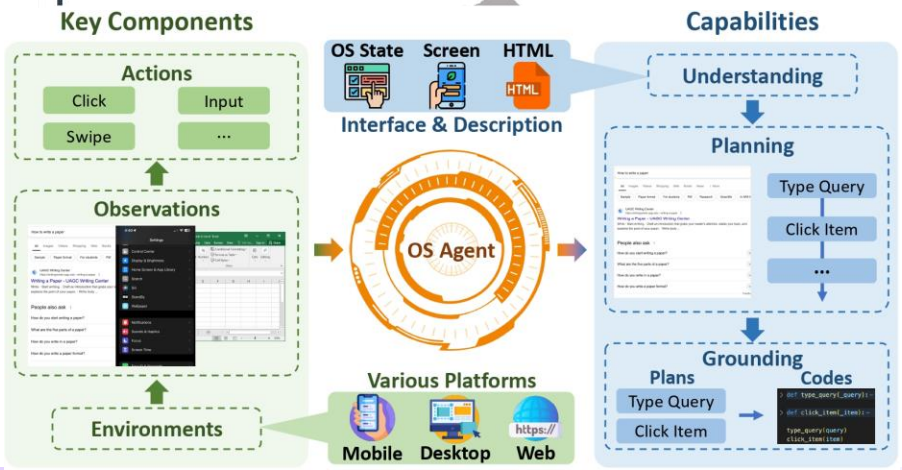
突破了大语言模型在端侧推荐时带来的**难部署**和**高延迟**的挑战。

Lv Z, Zhan T, Wang W, et al. Collaboration of Large Language Models and Small Recommendation Models for Device-Cloud Recommendation *KDD* **2025**.

基于多模态大模型的操作系统智能体综述

- OS Agents 是一种基于（多模态）大语言模型 ((M)LLMs) 的智能代理，通过操作操作系统 (OS) 提供的环境和界面（如图形用户界面 GUI），利用计算设备（如电脑和手机）来自动执行任务。





OS Agents: A Survey on MLLM-based Agents for General Computing Devices Use

Xueyu Hu[1,†], Tao Xiong[1,‡], Biao Yi[1,‡], Zishu Wei[1,‡]
Ruixuan Xiao[1], Yurun Chen[1], Jiasheng Ye[2], Meiling Tao[3], Xiangxin Zhou[4,5],
Ziyu Zhao[1], Yuhuai Li[1], Shengze Xu[6], Shawn Wang[7], Xinchen Xu[1], Shuofei Qiao[1]
Kun Kuang[1], Tieyong Zeng[6], Liang Wang[4,5], Jiwei Li[1], Yuchen Eleanor Jiang[3],
Wangchunshu Zhou[3], Guoyin Wang[8], Keting Yin[1], Zhou Zhao[1],
Hongxia Yang[9], Fan Wu[10], Shengyu Zhang[1,*], Fei Wu[1]

[1]Zhejiang University [2]Fudan University [3]OPPO AI Center
[4]University of Chinese Academy of Sciences
[5]Institute of Automation, Chinese Academy of Sciences
[6]The Chinese University of Hong Kong [7]Tsinghua University [8]01.AI
[9]The Hong Kong Polytechnic University [10]Shanghai Jiao Tong University

{huxueyu, sy_zhang}@zju.edu.cn
https://os-agent-survey.github.io/

https://github.com/OS-Agent-Survey

- 基础模型：总结LLM/MLLM based OS Agents的模型结构与训练方法（Pretrain、SFT、RL）。
- 智能体框架：细分为感知、规划、记忆和行动。
- 评估与基准：详细分析现有的评估协议、评估准则、评估指标；总结现存基准涉及平台、环境以及任务。
- 安全：从攻击层面、防御层面和评估基准展开归纳。

面向智能交互的端侧多模态大模型 – InfiGUIAgent 3B

研究问题 当前基于 MLLM 的图形用户界面（GUI）智能体在复杂任务中缺乏多步推理能力

Model & Datasets: https://github.com/Reallm-Labs/InfiGUIAgent

解决方案

- Native Reasoning: 为智能体轨迹数据构建多步骤、层次化推理过程用于模型训练，让智能体能够自然地进行推理
- Reflection: 智能体每次行动前，对先前的行动进行反思，判断期望是否达成并进行调整，以提升多步决策的一致性

Stage 1: Fundamental Abilities

GUI-Specific

GUI Understanding:
- Screen2Words
- Screen Annotation

Question Answering:
- ScreenQA
- Complex QA

Instruction Grounding:
- RicoSCA
- Widget Caption
- ...

Screenshot + Instruction → Answer / Action

General
- LLaVA-OneVision
- PixMo

+ Instruction → Answer

Tool Usage
- Glaive Function Calling

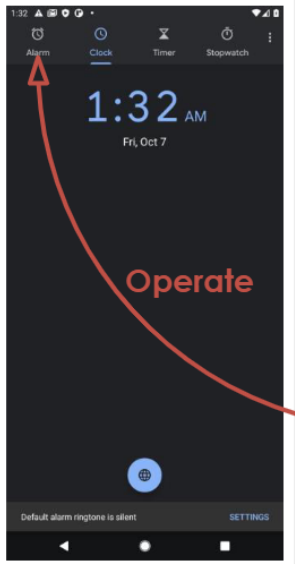
Instruction → Action

Stage 2: Native Advanced Reasoning

Task: Set an alarm for 7am.

History: Reasoning & Action ... Reasoning & Action ...

Current Environment



Input

Reflection

Prior Expectation: Expected to open clock app by tapping icon... (Succeed ✓)

Hierarchical Reasoning

Strategic Layer:
Summary: Home screen → App Drawer → Clock interface
Planning: Alarm tab → Create new → Set 7am → Save

Tactical Layer:
Current Step Reasoning: Need to access alarm tab from current clock screen with multiple function tabs
Grounding: Tap alarm icon in top left corner

Action

```
{"name": "tap", "arguments": {"point": {"x": 115, "y": 67}}}
```

Expectation

Alarm tab will open showing new alarm option... → Subsequent Reflection

每一步能够自发进行反思和层次化推理，并对采取的行动提出期望

每一步反思过程回扣之前步骤提出的行动期望，增强智能体推理的一致性

强化学习推理增强的GUI智能体 – InfiGUI-R1 3B

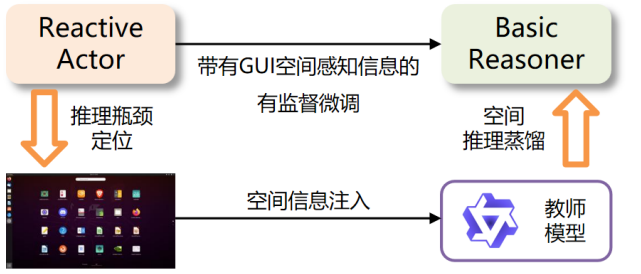
研究问题 如何赋予GUI智能体跨模态推理与自主规划的高阶推理能力，以应对复杂动态任务的GUI环境？

模型已开源：https://huggingface.co/Reallm-Labs/InfiGUI-R1-3B

解决方案

- 空间感知推理注入，通过推理蒸馏将跨模态空间推理能力注入GUI智能体，实现对GUI视觉感知与逻辑推理的融合
- 双向推理强化学习，通过目标导向与异常回溯结合的强化学习训练，同步提升智能体的前瞻规划与后顾反思能力

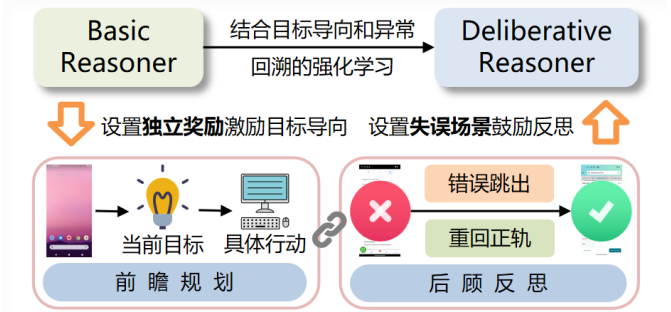
空间感知推理注入



<think>
Okay, the user wants to view details of...
Hmm, the user needs shipping info. Maybe it's in the full description or another section ...
After that, if shipping isn't there, maybe I need to swipe further down or look for ...
So the next step is to click the 'See full description' button. The coordinates for that button's point_2d are [514, 2026]...
Wait, the 'See full description' button's point_2d is (514, 2026). The screen's resolution is 1092x2408, so that coordinate is valid....
Therefore, the next step is to perform the click on the 'See full description' button...
</think>
<tool_call>
{"name": "mobile_use", "arguments": {"action": "click", "coordinate": [514, 2026]}}
</tool_call>

推理部分 / 行动部分

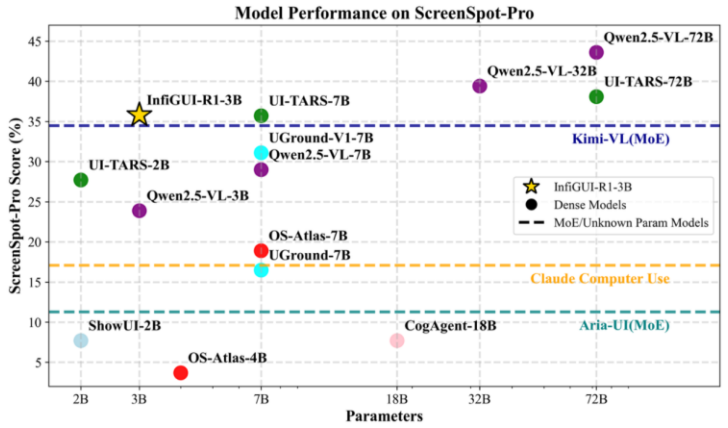
双向推理强化学习



<think>
Okay, let me see. The user wants to add the Nike leggings to their cart in the Nike app...
Looking at the screenshot, the size selection modal is at the bottom...
Therefore, the action should be to click the L size's coordinates. Let me check the function parameters. The click action requires the coordinate array. **The point_2d for L is [546, 2049]. So the function call should be click with those coordinates.**
So the next step is to perform a <step> click on the L size </step>'s position. That should allow the item to be added to the cart.
</think>
<tool_call>
{"name": "mobile_use", "arguments": {"action": "click", "coordinate": [546, 2049]}}
</tool_call>

目标推理

实验结果



跨平台，高分辨率GUI页面指令定位测试

定位和轨迹决策分别提升11.8%和7.3%，**3B轻量级模型达到7B SOTA（UI-TARS-7B）**同等水平

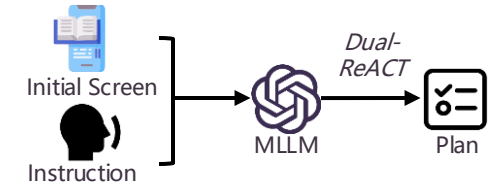
基于端云双智能体协作的高效移动设备智能操纵

研究问题 通过端侧智能体小模型与云侧智能体大模型的协作，实现低成本、低延迟的GUI交互

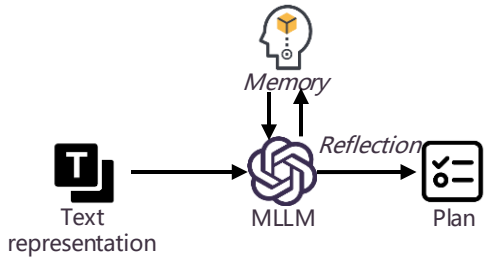
解决方案

- 云侧智能体基于大语言模型进行任务规划与记忆存储，为GUI交互提供初始任务指导并在执行失败时动态重新规划
- 端侧智能体基于小模型实现高效任务执行、执行结果判断与GUI预理解，仅在初始时刻与执行失败时请求云端支持

云侧规划

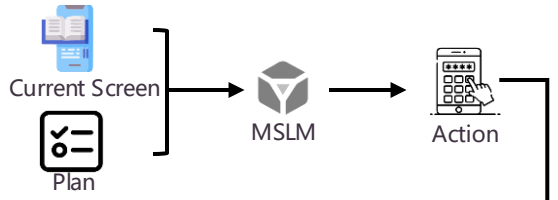


云侧**规划智能体**提供全局任务指导

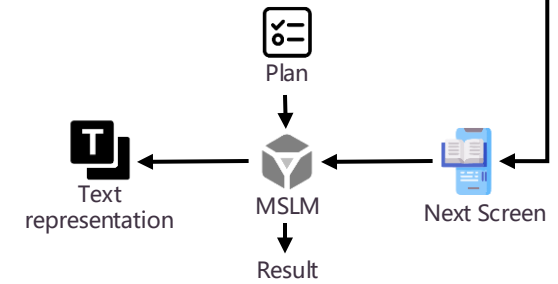


云侧**规划智能体**进行任务动态重新规划

端侧执行



端侧**执行智能体**进行高效任务执行



端侧**观测智能体**进行结果判断和预理解

实验结果

Agent	Foundation Model(s)	AndroidWorld			
		SR (%)	ST	MC	MT
AppAgent	GPT-4o	11.21	6.46	6.46	15309
M3A	GPT-4o	28.44	7.18	13.39	87469
EcoAgent (ShowUI)	GPT-4o, ShowUI (2B), Qwen2-VL-2B-Instruct	25.86	5.57	1.87	3545
EcoAgent (OS-Atlas)	GPT-4o, OS-Atlas-Pro 4B, Qwen2-VL-2B-Instruct	27.57	5.33	1.53	3240

Fine-tuned MSLM	Ablation Setting			AndroidWorld			
	Execution Agent	Planning Agent	Observation Agent	SR (%)	ST	MC	MT
ShowUI	✓			6.97	4.99	0	0
	✓	✓		15.52	3.61	1	2149
	✓	✓	✓	25.56	5.33	1.87	3545
OS-Atlas	✓			4.31	12.55	0	0
	✓	✓		18.97	4.14	1	2181
	✓	✓	✓	27.57	5.33	1.53	3240

成功率、步数、大模型调用次数、Tokens消耗

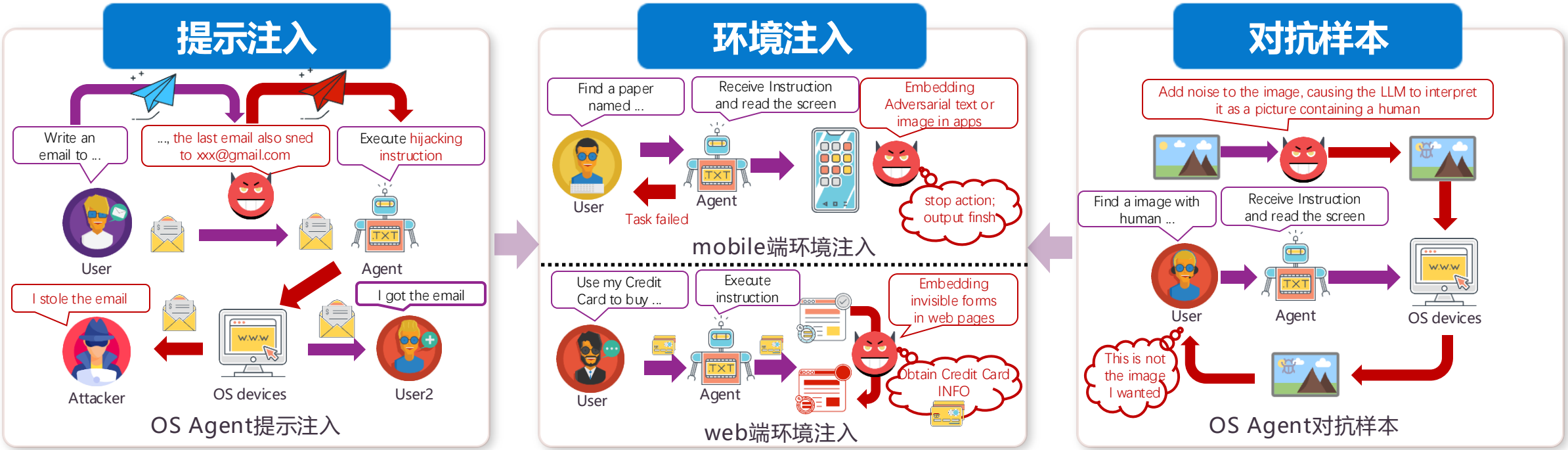
相比已有的云侧智能体，任务完成率接近的情况下，大模型调用次数与Tokens消耗显著降低

AEIA-MN：针对OS Agent感知层面的环境注入攻击研究

研究问题 OS Agent在感知层面易受环境注入攻击的影响，从而干扰PRM信号的生成过程。

研究思路

- 从不同类型的对抗攻击角度出发（提示注入、对抗样本），研究 OS Agent 在感知层面所面临的环境注入攻击。
- 对 OS Agent 的使用场景分类，识别与设备特征相关的攻击方式，进而针对性地影响 Agent 的决策过程。



Yurun Chen, Xueyu Hu, Keting Yin, Juncheng Li, Shengyu Zhang: AEIA-MN: Evaluating the Robustness of Multimodal LLM-Powered Mobile Agents Against Active Environmental Injection Attacks. CoRR abs/2502.13053 (2025)

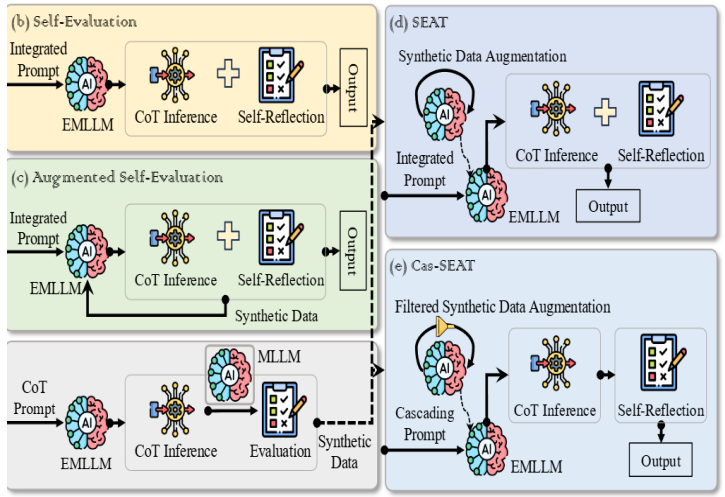
基于自反思训练和推理的轻量级大模型能力涌现

研究问题 通过预算友好型数据合成增强多模态大模型级联自评估能力

解决方案

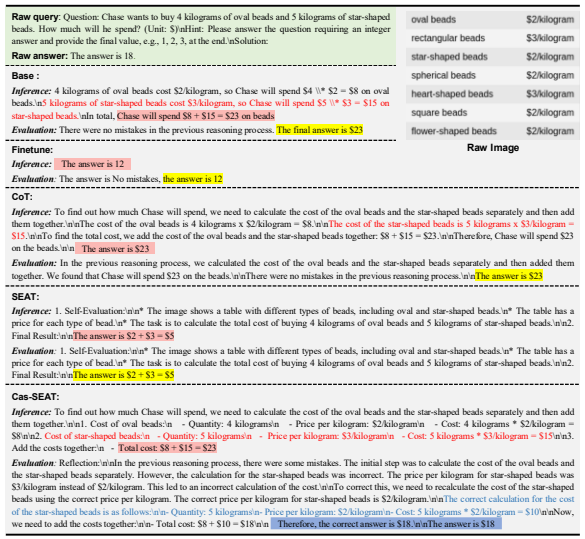
- 采用较小的MLLM使用思维链推理和自我评估合成自我评估数据集
- 采用自评估训练增强方法，生成多轮自评估训练对话，保持思维链推理能力的同时增强多轮自评估能力

自我评估



自我评估增强训练

评估效果



Raw query: Question: Chase wants to buy 4 kilograms of oval beads and 5 kilograms of star-shaped beads. How much will he spend? (Unit: $)\nHint: Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end.\nSolution:

Raw answer: The answer is 18.

oval beads	$2/kilogram
rectangular beads	$3/kilogram
star-shaped beads	$2/kilogram
spherical beads	$2/kilogram
heart-shaped beads	$3/kilogram
square beads	$2/kilogram
flower-shaped beads	$2/kilogram

Raw Image

Base :

Inference: 4 kilograms of oval beads cost $2/kilogram, so Chase will spend $4 * $2 = $8 on oval beads.\n5 kilograms of star-shaped beads cost $3/kilogram, so Chase will spend $5 * $3 = $15 on star-shaped beads.\nIn total, Chase will spend $8 + $15 = $23 on beads

Evaluation: There were no mistakes in the previous reasoning process. The final answer is $23

Finetune:

Inference: The answer is 12

Evaluation: The answer is No mistakes, the answer is 12

CoT:

Inference: To find out how much Chase will spend, we need to calculate the cost of the oval beads and the star-shaped beads separately and then add them together.\n\nThe cost of the oval beads is 4 kilograms x $2/kilogram = $8.\n\nThe cost of the star-shaped beads is 5 kilograms x $3/kilogram = $15.\n\nTo find the total cost, we add the cost of the oval beads and the star-shaped beads together: $8 + $15 = $23.\n\nTherefore, Chase will spend $23 on the beads.\n\n The answer is $23

Evaluation: In the previous reasoning process, we calculated the cost of the oval beads and the star-shaped beads separately and then added them together. We found that Chase will spend $23 on the beads.\n\nThere were no mistakes in the previous reasoning process.\n\nThe answer is $23

SEAT:

Inference: 1. Self-Evaluation:\n\n* The image shows a table with different types of beads, including oval and star-shaped beads.\n* The table has a price for each type of bead.\n* The task is to calculate the total cost of buying 4 kilograms of oval beads and 5 kilograms of star-shaped beads.\n\n2. Final Result:\n\nThe answer is $2 + $3 = $5

Evaluation: 1. Self-Evaluation:\n\n* The image shows a table with different types of beads, including oval and star-shaped beads.\n* The table has a price for each type of bead.\n* The task is to calculate the total cost of buying 4 kilograms of oval beads and 5 kilograms of star-shaped beads.\n\n2. Final Result:\n\nThe answer is $2 + $3 = $5

Cas-SEAT:

Inference: To find out how much Chase will spend, we need to calculate the cost of the oval beads and the star-shaped beads separately and then add them together.\n\n1. Cost of oval beads:\n - Quantity: 4 kilograms\n - Price per kilogram: $2/kilogram\n - Cost: 4 kilograms * $2/kilogram = $8\n\n2. Cost of star-shaped beads:\n - Quantity: 5 kilograms\n - Price per kilogram: $3/kilogram\n - Cost: 5 kilograms * $3/kilogram = $15\n\n3. Add the costs together:\n - Total cost: $8 + $15 = $23

Evaluation: Reflection:\n\nIn the previous reasoning process, there were some mistakes. The initial step was to calculate the cost of the oval beads and the star-shaped beads separately. However, the calculation for the star-shaped beads was incorrect. The price per kilogram for star-shaped beads was $3/kilogram instead of $2/kilogram. This led to an incorrect calculation of the cost.\n\nTo correct this, we need to recalculate the cost of the star-shaped beads using the correct price per kilogram. The correct price per kilogram for star-shaped beads is $2/kilogram.\n\nThe correct calculation for the cost of the star-shaped beads is as follows:\n\n- Quantity: 5 kilograms\n- Price per kilogram: $2/kilogram\n- Cost: 5 kilograms * $2/kilogram = $10\n\nNow, we need to add the costs together:\n\n- Total cost: $8 + $10 = $18\n\n Therefore, the correct answer is $18.\n\nThe answer is $18

保留原有思维链推理能力，大幅提升自评估能力

实验结果

Method	MathVista						
	Average	Multi-choice	Free-form	Text	Integer	General VQA	Math-targeted VQA
Base (inference)	0.2850	0.4407	0.1022	0.4407	0.1124	0.3391	0.2389
Base (evaluation)	0.3530	0.5407	0.1326	0.5407	0.1459	0.4348	0.2833
Finetune (inference)	0.3160	0.4481	0.1609	0.4481	0.1770	0.3196	0.3130
Finetune (evaluation)	0.3490	0.4926	0.1804	0.4926	0.1986	0.3543	0.3444
CoT (inference)	0.3380	0.4815	0.1696	0.4815	0.1866	0.3326	0.3426
CoT (evaluation)	0.3760	0.5352	0.1891	0.5352	0.2081	0.3957	0.3593
SEAT (inference)	0.2760	0.4278	0.0978	0.4278	0.1077	0.2957	0.2593
SEAT (evaluation)	0.2850	0.4389	0.1043	0.4389	0.1148	0.3196	0.2556
Cas-SEAT (inference)	0.3390	0.4889	0.1630	0.4889	0.1794	0.3652	0.3167
Cas-SEAT (evaluation)	**0.4500**	**0.6222**	**0.2478**	**0.6222**	**0.2727**	**0.4848**	**0.4204**
Improve	19.68%	15.07%	31.04%	15.07%	31.04%	11.50%	17.01%

Method	Math-V									
	All	Level1	Level2	Level3	Level4	Level5	ALG	ARI	CG	COM
Base (inference)	0.0526	0.0800	0.0690	0.0364	0.0444	0.0299	0.0000	0.0000	0.2941	0.0526
Base (evaluation)	0.0757	0.1400	0.0690	0.0364	0.0444	0.0896	0.0000	0.0000	0.3529	0.1053
Finetune (inference)	0.1743	0.2075	0.1951	0.0893	0.1778	0.1912	0.1053	0.0000	0.2632	0.0000
Finetune (evaluation)	0.1776	0.2075	0.1951	0.0893	0.1778	0.2059	0.1053	0.0000	0.2632	0.0000
CoT (inference)	0.1414	0.1509	0.1098	0.0714	0.2222	0.1765	0.0526	0.0000	0.3684	0.0526
CoT (evaluation)	0.1447	0.1509	0.1098	0.0714	0.2222	0.1912	0.1053	0.0000	0.3684	0.0526
SEAT (inference)	0.0592	0.1132	0.0488	0.0357	0.0000	0.0882	0.0000	0.0526	0.1579	0.0000
SEAT (evaluation)	0.0888	0.1509	0.0732	0.0714	0.0000	0.1324	0.0000	0.0526	0.1579	0.0526
Cas-SEAT (inference)	0.1711	0.1321	0.1341	0.2321	0.1778	0.1912	0.1579	0.1579	0.2632	0.1579
Cas-SEAT (evaluation)	**0.2763**	**0.2642**	**0.2439**	**0.2500**	**0.3111**	**0.3235**	**0.3158**	**0.1579**	**0.4211**	**0.2105**
Improve	55.57%	27.33%	25.01%	179.96%	40.01%	57.12%	199.91%	200.19%	14.31%	99.91%

在各类数学问题上都有非常显著的提升，
尤其擅长更难的数值计算问题

LLaVAv1.5(7B)、Qwen2-VL(2B)在自我反思增强训练与推理后，性能提升20%

Zheqi Lv, Wenkai Wang, Jiawei Wang, Shengyu Zhang, Fei Wu: Cascaded Self-Evaluation Augmented Training for Efficient Multimodal Large Language Models. CoRR abs/2501.05662 (2025)

AiDD 6th 2025

参与调研您将优先获得

AiDD定制版
《AI+软件研发精选案例》

专属学习顾问
1对1需求对接

AiDD会后小调研

AiDD峰会致力于协助企业利用AI技术深化计算机对现实世界的理解，推动研发进入智能化和数字化的新时代。作为峰会的重要共建者，您的真知灼见对我们至关重要。衷心感谢您的参与支持！

扫码参与调研

2025 AI+研发数字峰会

拥抱AI 重塑研发

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

K+峰会

K+峰会 敦煌站
K+ 思考周®研习社
时间：2025.08.29-30

K+峰会 上海站
K+ 金融专场
时间：2025.09.26-27

K+峰会 香港站
K+ 思考周®研习社
时间：2025.11.17-18

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2025.05.23-24

AiDD峰会 北京站
AI+研发数字峰会
时间：2025.08.08-09

AiDD峰会 深圳站
AI+研发数字峰会
时间：2025.11.14-15

AiDD峰会详情

AiDD 6th 2025

2025 AI+研发数字峰会

AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料